Submitted to the Annals of Applied Statistics

VARIABLE SELECTION FOR A CATEGORICAL VARYING-COEFFICIENT MODEL WITH IDENTIFICATIONS FOR DETERMINANTS OF BODY MASS INDEX

By Jiti Gao*, Bin Peng, Zhao Ren and Xiaohui Zhang

Monash University, University of Bath, University of Pittsburgh and University of Exeter

Obesity has become one of the major public health issues during the last three decades. A considerable number of determinants have been proposed for body mass index (BMI) by a large range of studies from multiple disciplines. In addition, it is well documented that impacts of these determinants are varying across demographic groups. However little is known about the relative importance of these potential determinants and the varying impacts of all relatively important determinants. Using the shrinkage estimation technique, we propose a variable selection procedure for the categorical varying-coefficient model. We present a simulation study to exam performance of our method in different scenarios. We further apply the proposed method to examine the impacts of a large number of potential determinants on BMI, using data from the 2013 National Health Interview Survey in the United States. By our method, the relevant determinants of BMI are identified through the variable selection procedure; and their varying impacts across demographic groups are quantified through the post-selection estimation.

1. Introduction. As a widely used measurement for body fat, body mass index (BMI) has been attracting significant attention from numerous researchers in multiple disciplines. The interest in measuring body fat came with increasing obesity in the last three decades especially in developed countries. According to WHO estimates, the worldwide prevalence of obesity is more than doubled between 1980 and 2014. Overweight is a major risk factor for a large range of noncommunicable diseases (Fontaine et al., 2003; WHO, 2015). It is thus crucial to identify and quantify the correlations between potential predictors and BMI. Empirical studies, which try to link particular lifestyle behaviours and other risk factors to BMI, may inform and guide policy makers to provide efficient incentives and interven-

*The first author acknowledges the financial support by the Australian Research Council Discovery Grants Program under Grant numbers: DP150101012 & DP170104421.

Keywords and phrases: Body Mass Index; Obesity; Optimal Variable Selection; Varying-Coefficient Regression

tions to reduce population BMI. Numerous studies have been seen in the last two decades and a large number of factors have been proposed as important drivers of increasing BMI (for references see Cawley (2011)). Though there is an impressive amount of evidence on the individual importance of determinants, there is little guidance for policy makers about where cost-containment efforts (Stice, Shaw and Marti, 2006) should be focused on. The inability of interventions to produce significant prevention effects may be due to incomplete understanding of the relative importance of predictors from various domains (Rehkopf et al., 2011).

A lot of effort has been devoted to selecting the relatively important predictors for BMI in the last decade. Besides the conventional, but controversial, stepwise regression procedures (e.g., Von Kries et al. (2002)), some new statistical methods have been proposed or adopted recently to select determinants of BMI. For example, Huang et al. (2009) propose a group bridge approach and apply it to determine risk factors on BMI of high school students. Rehkopf et al. (2011) adopt random forest, a tree-based analysis procedure, to rank the relative importance of risk factors for BMI among adolescent girls.

Despite the effort on selecting relatively important predictors for BMI, none of these studies simultaneously takes into account the fact that impacts of determinants on BMI may vary across demographic groups. In fact, these varying impacts have been well documented in the literature. For example, Yu (2012) find that education attainment has different impacts on BMI in different gender, age and race groups. Particularly, compared with college graduates, less educated whites and younger black women are more likely to be obese, and the differentials are larger for women than men, but weak or non-existent among black men and older black women. Similar evidences have been found by a considerable amount of studies, such as Colditz et al. (1991), Sobal, Rauschenbach and Frongillo (1992), Lipowicz, Gronkiewicz and Malina (2002), Zhang and Wang (2004), and so on. In order to capture such varying impacts, a common practice is to add interaction terms between selected BMI determinants and demographic variables into a regression model. The major shortcoming of this method is that it requires large degrees of freedom, which restricts the number of variables being allowed to have varying impacts on BMI. The choice of determinants having varying impacts, normally, serves to answer a specific research question, therefore it is arbitrary and lack of statistical support. Furthermore the method of adding interaction terms provides no statistical evidence to justify the importance of demographic variables, in terms of differencing the determinants' impacts on BMI.

In this paper, we provide a solution to the modelling issues existing in the literature of BMI studies using individual health survey data, i.e., (1) how to allow for and quantify the varying impacts of determinants on BMI; (2) how to justify the relative importance of demographic variables in differencing potential determinants' impacts on BMI; and (3) how to identify the relatively important determinants of BMI. Data used in this study are from 2013 National Health Interview Survey (NHIS) in the United States. There are 16,593 observations, 48 potential determinants and 32 demographic groups generated by 3 categorical variables (i.e., age group, gender and race).

To allow for and quantify the varying impacts of BMI determinants across demographic groups, we adopt the categorical varying-coefficient model proposed by Li, Ouyang and Racine (2013), which specifies the impacts of BMI determinants as unknown functions of demographic variables. Different from the conventional practice of adding interaction terms to regression models, categorical varying-coefficient model does not consume degrees of freedom that quickly when the number of demographic variables and/or BMI determinants increases.[1] Moreover, as documented in Li, Ouyang and Racine (2013), the selection of optimal bandwidths for categorical variables provides statistical justification on the relative importance of demographic variables in terms of differencing BMI determinants' impacts, and is able to serve as a filter to remove irrelevant demographic groups. For example, in our BMI study we are able to demonstrate that all demographic variables including age, gender and race are important in driving the BMI determinants' impacts to be different in different groups. We also find that gender and race are stronger in differencing the determinants' impacts on BMI than age. To identify the relatively important determinants of BMI, we adopt the group LASSO method proposed by Yuan and Lin (2006). In particular, we marry the categorical varying-coefficient model and the group LASSO method to simultaneously solve the aforementioned modelling issues in this BMI study.

The rest of the paper is organized as follows. We review the categorical varying-coefficient model of Li, Ouyang and Racine (2013), and introduce a variable selection procedure and its asymptotic results for the varying-coefficient model in Section 2. In Section 3, we conduct a Monte Carlo study to investigate the finite sample properties of the method. In Section 4, by using the 2013 NHIS data, we identify the important determinants of BMI and quantify their varying impacts on BMI across demographic groups. Section 5 concludes the paper with some discussions. The necessary assumptions required for the theoretical development are provided in Appendix A. Ad-

[1] A detailed example is provided in Appendix S3 of the supplementary file (Gao et al., 2017) to illustrate this difference.

ditional results and mathematical proofs are provided in the supplementary file of this paper (Gao et al., 2017).

2. Methodology. In this study, a categorical varying-coefficient model is adopted to capture the varying impacts of a large range of factors on BMI across demographic groups. Varying-coefficient models have attracted considerable attention and gained popularity in the past two decades from both theoretical and practical aspects (e.g., Hastie and Tibshirani, 1993; Fan and Zhang, 1999; Wang and Xia, 2009; Li and Racine, 2010; Li, Ouyang and Racine, 2013; and so forth). As discussed in Wang and Xia (2009), including spurious regressors can degrade the estimation efficiency substantially. In order to address this issue, variable selection for varying-coefficient models has received increasing attention (Wang, Li and Huang, 2008; Wang and Xia, 2009; Ma et al., 2015), but almost all of these existing variable selection methods for varying-coefficient models are specifically for the setting that only continuous predictors or indexes enter the nonparametric specification of linear parameters. In fact, it is very common in empirical applications that categorical variables influence the regressors' impacts on dependent variable, such as our BMI study in this paper.

To fill in the gap of literature and solve the modelling issues raised in BMI studies, we propose a variable selection procedure for the categorical varying-coefficient model below.

2.1. *Brief Review: A Categorical Varying-Coefficient Model.* The model of Li, Ouyang and Racine (2013) is specified as follows:

$$Y_i = X_i'\beta_0(Z_i) + \varepsilon_i, \quad i = 1, \ldots, N, \tag{2.1}$$

where $Z_i = (\bar{Z}_i', \tilde{Z}_i')'$ is an r-dimensional vector of discrete covariates with a support $\mathcal{D} = \bar{\mathcal{D}} \times \tilde{\mathcal{D}}$, $\bar{Z}_i = (Z_{i,1}, \ldots, Z_{i,\bar{r}})'$, $\tilde{Z}_i = (Z_{i,\bar{r}+1}, \ldots, Z_{i,r})'$ and $1 \leq \bar{r} \leq r$. Moreover, $\{\tilde{Z}_i, 1 \leq i \leq N\}$ is independent of all other variables and has no impact on $\beta_0(\cdot)$, which implies that $\tilde{Z}_i$ has no impact on Y_i at all. Therein, $\bar{Z}_i$ and $\tilde{Z}_i$ are referred to as relevant and irrelevant covariates respectively. When $\bar{r} = r$, there is no irrelevant covariate existing in the system, i.e., $\bar{Z}_i = Z_i$. To distinguish X_i from Z_i, they are referred to as regressors and covariates, respectively, hereafter. Based on the above description, the true model reduces to

$$Y_i = X_i'\beta_0(\bar{Z}_i) + \varepsilon_i, \quad i = 1, \ldots, N, \tag{2.2}$$

where ε_i is a random error term; $X_i = (X_{i,1}, \ldots, X_{i,p})'$ is a p-dimensional vector of regressors; $\beta_0(z) = (\beta_{01}(z), \ldots, \beta_{0p}(z))'$ is a p-dimensional unknown

coefficient function; and no information is known in advance to distinguish $\bar{Z}_i$ and $\tilde{Z}_i$. Moreover, both p and r are supposed to be fixed. This assumption is not that controversial. For example, in our BMI application, the sample size N is normally much larger than the number of potential predictors of X, i.e., p, and the number of possible covariates Z is even smaller. In particular, N, p and r are 16593, 48 and 3, respectively, in our BMI application. We refer to Section 4 for the details.

Applying model (2.2) to BMI data analysis allows us to capture the varying impacts of X, i.e., potential predictors such as lifestyles and socioeconomic factors, on BMI (indicated by Y) across demographic groups including gender, age group and race (denoted by Z). It is a common practice to capture such kind of varying impacts by adding interactions between the discrete Z variables and the X variables to a linear regression model, while it is straightforward to show that model (2.2) nests the latter model specification as a special case (cf., Appendix S3 of Gao et al. (2017)).

To carry on the regression, the kernel function of Aitchison and Aitken (1976) for an unordered covariate is adopted:

$$l(Z_{i,s}, z_s, \theta_s) = \begin{cases} 1, & \text{if } Z_{i,s} = z_s \\ \theta_s, & \text{otherwise} \end{cases}, \tag{2.3}$$

where the range of θ_s is $[0, 1]$ for $s = 1, \ldots, r$. It can be seen that $\theta_s = 0$ leads to an indicator function and $\theta_s = 1$ gives a uniform weight function. Then (2.3) allows us to construct a product kernel function of the form:

$$L(Z_i, z, \Theta) = \prod_{s=1}^{r} l(Z_{i,s}, z_s, \theta_s) = \prod_{s=1}^{r} \theta_s^{1(Z_{i,s} \neq z_s)}, \tag{2.4}$$

where $\Theta = (\theta_1, \ldots, \theta_r)'$. Therefore, for any $z \in \mathcal{D}$, the kernel-based OLS estimator is denoted as

$$\hat{\beta}(z) = \Big[\sum_{j=1}^{N} X_j X_j' L(Z_j, z, \hat{\Theta}) \Big]^{-1} \sum_{j=1}^{N} X_j Y_j L(Z_j, z, \hat{\Theta}),$$

where an optimal bandwidth $\hat{\Theta}$ is obtained by minimizing the following cross-validation criterion function:

$$CV(\Theta) = \frac{1}{N} \sum_{i=1}^{N} \left(Y_i - X_i' \hat{\beta}_{-i} \right)^2, \tag{2.5}$$

and the leave-one-out OLS estimator $\hat{\beta}_{-i}$ is defined as

$$\hat{\beta}_{-i} = \Big[\sum_{j=1, j\neq i}^{N} X_j X_j' L(Z_j, Z_i, \Theta) \Big]^{-1} \sum_{j=1, j\neq i}^{N} X_j Y_j L(Z_j, Z_i, \Theta).$$

It is convenient to introduce some notation here. For an r-dimensional vector $z = (z_1, \ldots, z_r)' \in \mathcal{D}$, we partition z as $z = (\bar{z}', \tilde{z}')'$ conformably with Z_i, where $\bar{z} = (z_1, \ldots, z_{\bar{r}})'$ and $\tilde{z} = (z_{\bar{r}+1}, \ldots, z_r)'$. Correspondingly, we partition Θ as $\Theta = (\bar{\Theta}', \tilde{\Theta}')'$, where $\bar{\Theta} = (\theta_1, \ldots, \theta_{\bar{r}})'$ and $\tilde{\Theta} = (\theta_{\bar{r}+1}, \ldots, \theta_r)'$. Due to space limitation, all assumptions needed for the lemmas and theorems in this paper are stated in the Appendix A, and all mathematical proofs are provided in the supplementary file (Gao et al., 2017). Given that our study is based on Li, Ouyang and Racine (2013), we borrow two results from them and summarize them in the following lemma:

LEMMA 2.1. *Let* $\hat{\Theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_r)' = \operatorname{argmin}_{\Theta \in [0,1]^p} CV(\Theta)$.

1. *Under Assumptions 1 and 2.1,* $\hat{\theta}_s = O_P\left(\frac{1}{N}\right)$ *for* $s = 1, \ldots, r$.
2. *Under Assumptions 1 and 2.2,* $\hat{\theta}_s = O_P\left(\frac{1}{\sqrt{N}}\right)$ *for* $s = 1, \ldots, \bar{r}$, *and* $\lim_{N\to\infty} \Pr(\hat{\theta}_{\bar{r}+1} = 1, \ldots, \hat{\theta}_r = 1) \geq \alpha$ *for some* $\alpha \in (0, 1)$.

Lemma 2.1 summaries Theorems 1 and 3 of Li, Ouyang and Racine (2013) and provides asymptotic theory of smoothing parameters $\hat{\Theta}$. In particular, the rate of convergence of $\hat{\theta}_s$ depends on whether there is irrelevant covariate or not, rather than the identification requirements stated in Assumption 2.1 or 2.2. For details, see Theorems 1 and 3 of Li, Ouyang and Racine (2013). It is worthwhile to mention that for nonparametric/varying-coefficient models with at least one covariate as continuous variable, the asymptotic theory of selected smoothing parameters through cross-validation has also been well developed (cf., Hall, Li and Racine (2007) and Li and Racine (2010)).

For a covariate z_s, if we obtain $\hat{\theta}_s = 1$, we can safely remove z_s from the model.[2] To some extent, this provides a variable selection procedure for covariates. Hereafter, with slight abuse of notation, we assume that we have removed all detected irrelevant covariates according to Lemma 2.1, i.e., those z_s with $\hat{\theta}_s = 1$, and the remaining covariates of the ith observation is still represented by $Z_i = (\bar{Z}_i', \tilde{Z}_i')'$ as before. However, clearly there is a positive

[2]Although one cannot always achieve $\hat{\theta}_s = 1$ for all irrelevant covariates simultaneously, as stated in Lemma 2.1, there is always a certain positive probability that we can recognize a covariate as irrelevant, i.e., the probability of $\hat{\theta}_s = 1$ for the corresponding covariate is positive.

probability such that no $\tilde{Z}_i$ exists. The purpose of this variable selection on covariates is to reduce the total number of distinct realizations of z from our sample $\{Z_1, \ldots, Z_N\}$.

2.2. *Variable Selection on X_i.* For model (2.2) with all detected irrelevant covariates removed, we propose a variable selection procedure to identify regressors of X_i with nonzero coefficient, when both p and r are fixed. Assume that there exists an unknown set $U^c \subseteq \{1, \ldots, p\}$ satisfying that $E|\beta_{0j}(\bar{Z}_i)|^2 = 0$ if and only if $j \in U^c$, where $\beta_{0j}(\bar{Z}_i)$ denotes the jth element of $\beta_0(\bar{Z}_i)$. To simplify notation, we assume that in the true model, $U = \{1, \ldots, p^*\}$ and $U^c = \{p^* + 1, \ldots, p\}$, where the integer p^* satisfies $1 \le p^* \le p$. In other words, only the first p^* variables in X_i have nonzero coefficients and our goal is to identify U and U^c.

Let m denote the number of realizations of z by observing $\{Z_1, \ldots, Z_N\}$. Obviously m converges to the cardinality of $\mathcal{D}$ in probability with nondegenerate probability imposed on i.i.d. Z_i as N diverges to ∞. Since m is finite and observable, our parameters of interest can be characterized by the following $m \times p$ matrix B with the underlying true coefficient function B_0. For the sake of presentation, denote

$$
\begin{aligned}
\underset{m\times p}{B} &= (\beta_1, \ldots, \beta_m)' = (b_1, \ldots, b_p), \\
\underset{p\times 1}{\beta_j} &= (\beta_{j,1}, \ldots, \beta_{j,p})' \text{ for } j = 1, \ldots, m, \\
\underset{m\times 1}{b_s} &= (\beta_{1,s}, \ldots, \beta_{m,s})' \text{ for } s = 1, \ldots, p, \\
\underset{m\times p}{B_0} &= (\beta_0(z^1), \ldots, \beta_0(z^m))' = (b_{01}, \ldots, b_{0p^*}, 0, \ldots, 0), \\
\underset{m\times 1}{b_{0s}} &= (\beta_{0s}(z^j), \ldots, \beta_{0s}(z^j))' \text{ for } s = 1, \ldots, p^*,
\end{aligned} \tag{2.6}
$$

where z^j, $j = 1, \ldots, m$, denotes the jth realization of $z \in \mathcal{D}$.

Notice that the last $p - p^*$ columns of B_0 are zero columns. By treating entries in each column of B_0 as a group, the selection on regressor of X_i is, essentially, to identify those groups (i.e., columns) of the matrix B_0 with all entries as zero. Following the spirit of Yuan and Lin (2006), we consider the following regularized least squares estimator:

$$
\hat{B} = (\hat{\beta}_{\gamma,1}, \ldots, \hat{\beta}_{\gamma,m})' = (\hat{b}_{\gamma,1}, \ldots, \hat{b}_{\gamma,p}) = \underset{B\in\mathbb{R}^{m\times p}}{\operatorname{argmin}} Q_\gamma(B), \tag{2.7}
$$

and

$$
Q_\gamma(B) = \sum_{j=1}^{m} \sum_{i=1}^{N} \left(Y_i - X_i'\beta_j\right)^2 L(Z_i, z^j, \hat{\Theta}) + \sum_{s=1}^{p} \gamma_s \|b_s\|, \tag{2.8}
$$

where $\hat{\Theta}$ is the smoothing parameter vector obtained from Lemma 2.1; b_s $(s = 1, \ldots, p)$ is the sth column of B as denoted in (2.6); $\sum_{s=1}^{p} \gamma_s \|b_s\|$ is the group-wise regularizer and defined as the weighted sum of the ℓ_2 norms of all the column vectors in B; and $\gamma = (\gamma_1, \ldots, \gamma_p)'$ represents the weight that controls the group-wise regularizer.

REMARK 2.1. *If we ignore the optimal bandwidth selection and use an indicator function to replace all kernel functions, we essentially have an adaptive version of a group LASSO model (cf., Yuan and Lin (2006)). On the other hand, if we set all γ_s's to 0, we end up with the model proposed in Li, Ouyang and Racine (2013). Due to the features of BMI data, we combine both methods together and try to filter out any redundant information as much as possible.*

Our first theorem is stated below.

THEOREM 2.1. *Suppose Assumptions 1-3 hold.*

1. *Let $\gamma^* = (\gamma_1, \ldots, \gamma_{p^*})'$ and $\frac{\|\gamma^*\|}{\sqrt{N}} \to \omega_1$, where ω_1 is a constant satisfying $0 \le \omega_1 < \infty$. Then $\left\|\hat{\beta}_{\gamma,j} - \beta_0(\bar{z}^j)\right\| = O_P\left(N^{-1/2}\right)$ for $j = 1, \ldots, m$, where $\bar{z}^j = (z_1^j, \ldots, z_{\bar{r}}^j)'$.*
2. *Let $\frac{1}{\sqrt{N}} \min_{s \in \{p^*+1, \ldots, p\}} \gamma_s \ge \omega_2$, where ω_2 is a sufficiently large constant. Then $\Pr(\|\hat{b}_{\gamma,j}\| = 0) \to 1$ for $j = p^* + 1, \ldots, p$.*

The first result of Theorem 2.1 states if the regularizer weight is not too large, estimator (2.7) always has optimal $\sqrt{N}$ consistency. The second result implies that when the regularizer weight is at level $\sqrt{N}$, estimator (2.7) can successfully identify those regressors with zero coefficient. To satisfy the assumptions in Theorem 2.1, all elements of γ can be simply set at level $\sqrt{N}$. However, with such γ, Theorem 2.1 does not imply any asymptotic normality property of the estimator (2.7). While in Li, Ouyang and Racine (2013), asymptotic normality property has been achieved for the oracle estimator.[3] Specifically, the oracle estimator is defined as:

$$\hat{\beta}_{ora}(\bar{z}^j) = \left(\sum_{i=1}^{N} X_{iU} X_{iU}' L(Z_i, z^j, \hat{\Theta})\right)^{-1} \sum_{i=1}^{N} X_{iU} Y_i L(Z_i, z^j, \hat{\Theta}), \tag{2.9}$$

[3]Notice that the word "oracle" refers to those estimators provided in Li, Ouyang and Racine (2013) by assuming we know the true set U. Here we completely ignore the inefficiency brought in the model by the irrelevant covariates $\tilde{Z}_i$. The asymptotically efficient estimator is obtained when we know both the set U and the irrelevant covariates. However, this can only be done at certain probability based on Lemma 2.1.

where $j = 1, \ldots, m$ and $X_{iU} = (X_{i,1}, \ldots, X_{i,p^*})'$.

In fact, with a more careful data-driven choice of γ, we can further achieve the asymptotic normality whenever there is no irrelevant covariate with the help of following oracle property for our estimator (2.7).

THEOREM 2.2. *Under conditions of Theorem 2.1,* $\left\|\hat{\beta}_{\gamma,jU} - \hat{\beta}_{ora}(\bar{z}^j)\right\| = O_P\left(\frac{\|\gamma^*\|}{N}\right)$ *for* $j = 1, \ldots, m$, *where* $\hat{\beta}_{\gamma,jU} = (\hat{\beta}_{\gamma,j1}, \ldots, \hat{\beta}_{\gamma,jp^*})'$; $\hat{\beta}_{\gamma,js}$ *denotes the* s*th element of* $\hat{\beta}_{\gamma,j}$ *for* $j = 1, \ldots, m$ *and* $s = 1, \ldots, p^*$; *and* γ^* *is denoted in Theorem 2.1.*

To achieve an asymptotic normality for the estimator (2.7), the convergence rate of $\hat{\beta}_{\gamma,jU}$ to $\hat{\beta}_{ora}(\bar{z}^j)$ has to be much faster than $\frac{1}{\sqrt{N}}$. The oracle property in Theorem 2.2 implies such a result as long as $\|\gamma^*\|$ is much smaller than $\sqrt{N}$. Therefore the simple choice of $\sqrt{N}$ level for γ is not sufficient.

To achieve a desired asymptotic normality property for the estimator (2.7), we propose a data-driven choice of γ, which can yield an even faster rate of convergence of an order of $o_P\left(\frac{1}{\sqrt{N}}\right)$ to the oracle estimator. From now on, we assume that whenever the true coefficient is nonzero, that is $b_{0s} \neq 0$ for $s = 1, \ldots, p^*$, its ℓ_2 norm is much larger than root N level, i.e., $\|b_{0s}\| \gg \frac{1}{\sqrt{N}}$ for $s = 1, \ldots, p^*$. This assumption is not controversial in the current fixed dimension setting in which $\|b_{0s}\|$ is some positive constant as N increases.

Similarly to Wang and Leng (2007) and Wang and Xia (2009), our data-driven regularizer weight is as follows:

$$\gamma = \tilde{\gamma}\left(\|\tilde{b}_1\|^{-1}, \ldots, \|\tilde{b}_p\|^{-1}\right)', \tag{2.10}$$

where $\tilde{\gamma}$ is a scalar, $\tilde{b}_s$ is the sth column of the unregularized estimator $\tilde{B}$, and $\tilde{B}$ is obtained from (2.8) by simply choosing $\gamma_1 = \cdots = \gamma_p = 0$ as follows:

$$\tilde{B} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_m)' = (\tilde{b}_1, \ldots, \tilde{b}_p) = \underset{B \in \mathbb{R}^{m \times p}}{\operatorname{argmin}} Q(B) \tag{2.11}$$

and

$$Q(B) = \sum_{j=1}^{m} \sum_{i=1}^{N} \left(Y_i - X_i' \beta_j\right)^2 L(Z_i, z^j, \hat{\Theta}). \tag{2.12}$$

Under Assumption 3.1, the first result of Theorem 2.1 and the assumption of $\|b_{0s}\| \gg \frac{1}{\sqrt{N}}$ for $s = 1, \ldots, p^*$, it is easy to verify that $\|\tilde{b}_s\|^{-1} = o_P(\sqrt{N})$

for $s = 1, \ldots, p^*$ and $\|\tilde{b}_s\| = O_P(1/\sqrt{N})$ for $s = p^* + 1, \ldots, p$. Then the intuition of choosing γ as (2.10) is straightforward. The unregularized estimator $\tilde{B}$ is an $\sqrt{N}$ consistent estimator. It provides information on how likely each column of B_0 is a zero column. In other words, smaller $\|\tilde{b}_j\|$ implies that the jth column is more likely to be zero and hence suggests a larger regularizer on $\|b_j\|$. In particular, given that $\|\tilde{b}_s\|^{-1} = o_P(\sqrt{N})$ for $s = 1, \ldots, p^*$, Theorem 2.2 implies the desired rate of $o_P\left(\frac{1}{\sqrt{N}}\right)$ for $\hat{\beta}_{\gamma,jU}$ to be the oracle estimator $\hat{\beta}_{ora}(\bar{z}^j)$. Given the form of γ in (2.10), the selection on the vector γ reduces to the selection on the scalar $\tilde{\gamma}$. Note that the properties of $\|\tilde{b}_j\|^{-1}$ for $j = 1, \ldots, p$ imply that a large enough constant $\tilde{\gamma}$ would satisfy all the conditions on γ. More specifically, we select the constant $\tilde{\gamma}$ by the following modified BIC-type (MBIC) criterion:

$$MBIC_{\tilde{\gamma}} = \ln RSS_{\tilde{\gamma}} + df_{\tilde{\gamma}} \cdot \frac{\ln N}{N},$$

where $df_{\tilde{\gamma}}$ is the number of nonzero coefficients identified by $\hat{B}_{\tilde{\gamma}}$, and $RSS_{\tilde{\gamma}}$ is defined as $RSS_{\tilde{\gamma}} = \frac{1}{N}\sum_{j=1}^{m}\sum_{i=1}^{N}\left(Y_i - X_i'\hat{\beta}_{\tilde{\gamma},j}\right)^2 L(Z_i, z^j, \hat{\Theta})$. The weight parameter is obtained by

$$\hat{\tilde{\gamma}} = \operatorname*{argmin}_{\tilde{\gamma}} MBIC_{\tilde{\gamma}}. \tag{2.13}$$

Recall the true set of nonzero coefficients is denoted by $U = \{1, \ldots, p^*\}$. Let $S_{\hat{\tilde{\gamma}}} = \{j : \|\hat{b}_{\hat{\tilde{\gamma}},j}\| > 0, 1 \le j \le p\}$ indicate the set of relevant variables identified by the regularized estimator $\hat{B}_{\hat{\tilde{\gamma}}}$ with the weight parameter $\hat{\tilde{\gamma}}$ chosen by (2.13). Then we have

THEOREM 2.3. *Suppose that $\|b_{0s}\| \gg \frac{1}{\sqrt{N}}$ for $s = 1, \ldots, p^*$. Under conditions of Theorem 2.1, the weight parameter selected by the modified BIC-type criterion (2.13) can:*

1. *Identify the true model consistently, i.e., $\Pr(S_{\hat{\tilde{\gamma}}} = U) \to 1$ as $N \to \infty$;*
2. *Achieve asymptotic normality, i.e.,*

$$\sqrt{N}(\hat{\beta}_{\hat{\tilde{\gamma}},jU} - \beta_{0U}(z^j)) \to_D N(0, \Sigma(z^j)) \tag{2.14}$$

for the relevant covariate case defined in Assumption 2, and for $j = 1, \ldots, m$, where

$$\Sigma(z^j) = A^{-1}(z^j)\Omega(z^j)A^{-1}(z^j), \quad A(z^j) = E[X_{iU}X_{iU}'|z^j]\Pr(z^j),$$
$$\Omega(z^j) = E[\varepsilon_i^2 X_{iU}X_{iU}'|z^j]\Pr(z^j), \quad \beta_{0U}(z^j) = (\beta_{01}(z^j), \ldots, \beta_{0p^*}(z^j))',$$

and X_{iU} has been defined in (2.9).

3. For irrelevant covariate case defined in Assumption 2,

$$\hat{\beta}_{\hat{\bar{\gamma}},jU} - \beta_{0U}(\bar{z}^j) = O_P\left(\frac{1}{\sqrt{N}}\right) \tag{2.15}$$

for $j = 1, \ldots, m$, *where* $\beta_{0U}(\bar{z}^j) = (\beta_{01}(\bar{z}^j), \ldots, \beta_{0p^*}(\bar{z}^j))'$.

When there is no irrelevant covariate (i.e., $r = \bar{r}$ and $Z_i = \bar{Z}_i$), the asymptotic normality result of (2.14) is based on the limiting distribution of $\sqrt{N}(\hat{\beta}_{ora}(z^j) - \beta_{0U}(z^j))$, which is established by applying Theorem 2 of Li, Ouyang and Racine (2013) on the oracle model. In practice, one may want to establish a consistent estimate for $\Sigma(z^j)$ for $j = 1, \ldots, m$, which can be immediately obtained following the procedure provided in Theorem 2 of Li, Ouyang and Racine (2013), assuming $S_{\hat{\bar{\gamma}}} = U$:

$$\hat{\Sigma}(z^j) = \hat{A}^{-1}(z^j)\hat{\Omega}(z^j)\hat{A}^{-1}(z^j),$$

where $\hat{\varepsilon}_i = Y_i - X_i'\hat{\beta}_{\hat{\bar{\gamma}},jU}$, $\hat{\Omega}^{-1}(z^j) = \frac{1}{N}\sum_{i=1}^N \hat{\varepsilon}_i^2 X_{iU}X_{iU}'L(Z_i, z^j, \hat{\Theta})$, and $\hat{A}^{-1}(z^j) = \frac{1}{N}\sum_{i=1}^N X_{iU}X_{iU}'L(Z_i, z^j, \hat{\Theta})$.

However, when there are irrelevant covariates (i.e., $r > \bar{r}$), the asymptotic distribution of $\sqrt{N}(\hat{\beta}_{ora}(\bar{z}^j) - \beta_{0U}(\bar{z}^j))$ remains unknown even for the oracle estimator and hence we only obtain $\sqrt{N}$ consistency in (2.15). In this case, the asymptotic distribution of $\sqrt{N}(\hat{\beta}_{ora}(\bar{z}^j) - \beta_{0U}(\bar{z}^j))$ can be established by using a bootstrap method as documented in Li, Ouyang and Racine (2013).

In this section, we propose a regularized estimator for the categorical varying-coefficient model and obtain its superior statistical properties. In particular, the coefficients of the proposed categorical varying-coefficient model possess a natural group structure. To take an advantage of the structure, we apply a group-wise regularizer to improve accuracy of variable selection and parameter estimation. Moreover, we apply a data-driven method, i.e., a modified BIC-type criterion, to select the weight parameter, which further boosts the performance and helps to achieve an asymptotic normality property for the estimator, especially when no irrelevant covariate presents.

3. Monte Carlo Evidence. In this section, we conduct a comprehensive Monte Carlo (MC) study to show the finite-sample performance of our method and a range of competing methods. To each generated data set $\{Y_i, X_i, Z_i\}$, firstly, we apply model (2.2) and estimate the optimal bandwidths. Following Lemma 2.1 and its discussion in Section 2.1, we remove irrelevant covariates to reduce the number of groups based on the realizations of Z_i.[4] Secondly, we identify the irrelevant regressors by estimating $\hat{B}$

[4]Refer to Li, Ouyang and Racine (2013) for extensive evidences on the performance of bandwidth selection in finite sample.

through (2.7). Lastly, we estimate the model excluding irrelevant covariates and regressors by the unregularized estimator proposed in Li, Ouyang and Racine (2013). The purpose of the last step is to further reduce the possible bias.

To compare the finite-sample performance of our method with some competing ones and put all the methods on equal footing, we use their adaptive versions for all LASSO related methods. More specifically, for each data set, we conduct (a) an adaptive version group LASSO estimation method; (b) an adaptive version of LASSO estimation method; and (c) stepwise estimation method. In particular, group LASSO method (denoted by GroupL) is essentially a special case of (2.2), i.e., with all bandwidths equal to 0. Alternatively, without taking into account of the varying impacts of X on Y according to Z, we apply methods (b) and (c) to the linear regression model (3.1) below (denoted by LASSO1 and SW1, respectively). Moreover, we apply methods (b) and (c) to the linear regression model (3.2) below (denoted by LASSO2 and SW2, respectively), where the varying impacts of X on Y are (particularly) captured by the interaction terms between X on Z. It is a very common practice in empirical studies (e.g., Yu (2012)).

$$Y_i = \left(X_i', Z_{i,11}, \ldots, Z_{i,1c_1-1}, \ldots, Z_{i,r1}, \ldots, Z_{i,rc_r-1}\right)' \beta_0^* + \varepsilon_i, \tag{3.1}$$

$$\begin{aligned} Y_i = \Big(&X_i', (Z_{i,11}X_i)', \cdots, (Z_{i,1c_1-1}X_i)', \\ &(Z_{i,r1}X_i)', \cdots, (Z_{i,rc_r-1}X_i)'\Big)' \beta_0^* + \varepsilon_i, \end{aligned} \tag{3.2}$$

where $Z_{i,jk} = 1$ if the j^{th} element of Z_i being k with $k = 1, \ldots, c_j - 1$; $Z_{i,jk} = 0$, otherwise.

Notice that when X_i does not exist in a model (3.1), i.e., only categorical variables are included, special treatment (Gertheiss and Tutz, 2010) can be considered. We avoid using more complicated ways to introduce interactions in model (3.2), since it is almost impossible to exhaust all possibilities.

We consider three scenarios in terms of the data generating process (DGP). In the first two scenarios, the DGPs are based on two categorical varying-coefficient models, i.e., without and with irrelevant covariate included in Z_i, respectively. And the DGP of the third scenario is a conventional linear regression model. Details of the DGPs are as follows.

Scenario 1: Let $p = 10$, $p^* = 5$, and $Y_i = (1,\ X_i')'\beta_0(Z_i) + \varepsilon_i$, where $X_i = H_i + V_i$ and $Z_i = (Z_{i,1}, \cdots, Z_{i,r})'$. For $\forall j = 1, \ldots, r$, $Z_{i,j}$ is i.i.d. over i and takes a value from $\{0, 1, 2\}$ with probability $\{0.25, 0.25, 0.5\}$ respectively. V_i is i.i.d. over i and follows $N(Z_{i,1}/2 \cdot i_{p-1}, \sqrt{Z_{i,1}+1} \cdot I_{p-1})$, in which I_{p-1} denotes the $(p-1)$-dimensional identity matrix and i_{p-1} represents the

$(p-1)$-dimensional vector with all entries being one; H_i is i.i.d. over i and follows $N(i_{p-1}, I_{p-1})$; and ε_i is i.i.d. over i and follows $N(0,1)$. Let $\beta_{0j}(Z_i)$ denote the jth element of the coefficient function $\beta_0(Z_i)$ for $j = 1, \ldots, p$.

Two sub-scenarios are designed as without and with irrelevant covariate included in Z_i, respectively.

- **Scenario 1.1:** Relevant Covariate Case (i.e., $\bar{r} = r$). For $\forall j \leq 5$,

$$\beta_{0j}(Z_i) = \begin{cases} 2+2j, & \text{if the remainder of } \sum_{k=1}^{r} Z_{i,k}/2 \text{ is } 0 \\ 1+2j, & \text{otherwise} \end{cases};$$

  for $\forall j > 5$, $\beta_{0j} = 0$.
- **Scenario 1.2:** Irrelevant Covariate Case (i.e., $\bar{r} = 1$). For $\forall j \leq 5$,

$$\beta_{0j}(Z_i) = \begin{cases} 2+2j, & \text{if the remainder of } Z_{i,1}/2 \text{ is } 0 \\ 1+2j, & \text{otherwise} \end{cases};$$

  for $j > 5$, $\beta_{0j} = 0$.

Scenario 2: Let $Y_i = (1,\ X_i')'\beta_0 + \varepsilon_i$, where $\beta_0 = (\beta_{01}, \ldots, \beta_{0p})'$, and $\beta_{0j} = 5$ with $j \leq 5$ and $\beta_{0j} = 0$ with $j > 5$. All the other variables are generated in exactly the same way as for Scenario 1.

Under Scenario 1, model (2.2) is correctly specified, while models (3.1) and (3.2) are misspecified. Therefore, we expect our estimator performs better than the other methods. Under Scenario 2, all models (i.e., (2.2), (3.1) and (3.2)) are correctly specified, so we expect reasonable performance from all the estimators.

To evaluate model performance, we examine three measures. They are (1) the percentage of missed true regressors (FNR); (2) the percentage of falsely selected noise regressors (FPR)[5]; and (3) mean squared prediction error (MSPE). We calculate MSPE, in the spirit of Chu, Li and Reimherr (2016), as follows:

$$MSPE = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_{-i} - y_i)^2, \tag{3.3}$$

where $\hat{y}_{-i}$ denotes the leave-one-out prediction for the ith individual (i.e., we implement estimation without the observation of the ith individual, and then use the estimated parameters to predict y_i for the ith individual). For

[5]To be clear, all binary variables and interactions terms in (3.1) and (3.2) are considered as redundant information. For example, if we identify some interaction terms as relevant regressors by LASSO method for model (3.2), these variables are counted as falsely selected.

each method under each scenario, we report averaged, over 1000 replications, FNR and FPR, and the root of averaged MSPE, denoted as RME. Note that the estimated RME should ideally converge to the standard deviation of ε_i (i.e., 1 in our MC design). Therefore, an estimated RME closing to 1 is an indicator for good model performance of the corresponding method.

In this MC study, we also consider a range of different settings for (N, r). In particular we consider N of 2000, 4000 and 8000, which are reasonable, if not much smaller, sample sizes in empirical applications. With regard to the size of r, we set it as 2, 3, and 4. It is noteworthy that as $r = 4$, we already have 81 demographic groups based on our DGP, so it is more than enough to demonstrate that the current setting covers our case study perfectly. For example, in our BMI study, 3 covariates (and 32 groups) are reasonably considered, which is supported by the BMI literature (cf., Yu (2012)).

We summarize the simulation results in Table 1. As expected, under Scenarios 1.1 and 1.2, our estimator (denoted as Varying-Coef) and group LASSO estimator (denoted as GroupL) outperform all other methods in general. As models (3.1) and (3.2) are misspecified, it is not surprising that LASSO1, LASSO2, SW1 and SW2 do not perform well. The RME's estimated by our estimator and group LASSO method, under different settings, are all close to 1, i.e., the true standard deviation of ε_i. However, those estimated by LASSO and stepwise methods are far away from 1, which is an indication for less accurate estimates. Note that the true regressor can almost be identified by our estimator and group LASSO method, i.e., FNR's are zero; in contrast, FNR's from SW1, SW2 and LASSO2 are considerably large. FPR's from Varying-Coef and GroupL are very low, compared to those from all other methods. Not surprisingly, under Scenario 2, all methods perform relatively well except SW1 and SW2.

We now take a close look at these results from Varying-Coef and GroupL, as both of them can address two questions raised in the introduction, i.e., (1) allowing for and quantifying the varying impacts, and (2) identifying the relatively important determinants. However, only our method is able to address the question of "how to justify the relative importance of demographic variables" by looking at the estimates of the optimal bandwidths based on Lemma 2.1. Compared to the group LASSO method, the better performance of the varying-coefficient setting is due to the following two reasons: (1) The varying-coefficient setting uses optimal bandwidths throughout Scenarios 1.1, 1.2 and 2, so the RMEs of Varying-Coef are closer to 1 as expected; and (2) For Scenario 1.2, the varying-coefficient setting can potentially throw away more possible irrelevant variables, so that reduces the number of groups based on the realizations of Z_i. In other words, each group

can potentially include more samples after we remove extra covariates from the system. For the sake of space, we report the histograms of the estimates on the bandwidth of irrelevant covariate with corresponding discussions in the supplementary file of this paper (Gao et al., 2017).

4. An Application to BMI.

4.1. *Data.* Data used in this empirical study are from the 2013 National Health Interview Survey (NHIS) in the United States. The NHIS is conducted annually through face-to-face interviews. Our analysis focuses on adults aged 18 and over. BMI is calculated based on self-reported height and weight. We exclude underweight individuals (BMI less than 18.5) from our analysis, and focus on such individuals with normal weight and overweight. There are three reasons for us to do so. First, underweight is a much less prevalent health problem in developed countries like the U.S. In particular, in the NHIS data underweight accounts for a very small proportion, i.e., 1.8 percent of the whole sample. Second, factors causing (or relating to) underweight are very much different from those for overweight or obesity. For example, eating disorders, such as anorexia nervosa and bulimia, lack of nutrition, and a hypermetabolism state, are considered as causes of underweight (Ali and Lindström, 2006), while unhealthy lifestyles and poor socio-economic factors are the major determinants of overweight and obesity (as discussed below in detail). However, information on these potential determinants of underweight is not available in NHIS. Last but not least, for common factors causing both underweight and overweight, their impacts on BMI might have different signs. For example, mental health problems, such as depression, can cause both BMI increase from normal weight to overweight level (positive impact on BMI) (Faith et al., 2011) and BMI decrease from normal weight to underweight level (negative impact) (Carey et al., 2014). This kind of "U" shape impact of determinants on BMI is hardly captured by our method.[6] In the end we use the natural logarithm transformed BMI in our analysis, because BMI scores are skewed towards higher values in our sample (Zeng et al., 2013).

Through a systematic review of the literature on overweight and obesity, we test impacts of 48 factors[7] (i.e., regressors X in the model (2.2)) on BMI,

[6]We thank one referee for pointing out that quantile regression can serve as an alternative modelling method for BMI (Koenker, 2005; Zhao, Zhang and Liu, 2014). See Section 5 for a detailed discussion.

[7]The number of factors tested is restricted by information available in the data set. For example, energy intake and dietary habit are important factor for BMI and obesity (see, for example, Hill and Peters (1998)). But information about food consumption is not

TABLE 1. *Monte Carlo Simulation Results*

			Varying-Coef			GroupL			LASSO1			SW1			LASSO2			SW2		
	r	N	RME	FNR	FPR	RME	FNR	FPR	RME	FNR	FPR	RME	FNR	FPR	RME	FNR	FPR	RME	FNR	FPR
Scenario 1.1	2	2000	0.9871	0.0000	0.0357	0.9869	0.0000	0.0381	4.1390	0.0000	0.2497	4.2554	0.0833	0.1843	3.4135	0.0158	0.6561	3.4344	0.0143	0.1852
		4000	0.9942	0.0000	0.0076	0.9941	0.0000	0.0078	4.1424	0.0000	0.2168	4.2504	0.0833	0.1888	3.3459	0.0166	0.6567	3.4270	0.0143	0.2081
		8000	0.9970	0.0000	0.0031	0.9966	0.0000	0.0036	4.1458	0.0000	0.1878	4.2528	0.0833	0.1889	3.2587	0.0170	0.6567	3.3922	0.0143	0.2406
	3	2000	0.9354	0.0000	0.0404	0.9321	0.0000	0.0445	4.2912	0.0000	0.2682	4.4940	0.0769	0.1192	4.1787	0.0160	0.6589	4.4234	0.0121	0.1145
		4000	0.9801	0.0000	0.0118	0.9794	0.0000	0.0149	4.2993	0.0000	0.2160	4.4440	0.0769	0.1827	4.1822	0.0164	0.6574	4.3875	0.0121	0.1459
		8000	0.9909	0.0000	0.0068	0.9907	0.0000	0.0075	4.3031	0.0000	0.2401	4.4092	0.0769	0.2376	5.9979	0.0165	0.6577	4.3494	0.0121	0.1729
	4	2000	0.8038	0.0000	0.0921	0.7565	0.0000	0.0934	4.3319	0.0000	0.2264	4.6127	0.0714	0.0583	4.5710	0.0163	0.6583	4.6233	0.0105	0.0684
		4000	0.9585	0.0000	0.0758	0.8932	0.0000	0.0802	4.3393	0.0000	0.1693	4.5909	0.0714	0.0812	4.5658	0.0158	0.6565	4.6252	0.0105	0.0854
		8000	0.9986	0.0000	0.0660	0.9477	0.0000	0.0690	4.3433	0.0000	0.1602	4.5340	0.0714	0.1375	4.2906	0.0162	0.6561	4.6150	0.0105	0.1123
Scenario 1.2	2	2000	0.9929	0.0000	0.0379	0.9868	0.0000	0.1639	3.4909	0.0000	0.1383	3.6608	0.0833	0.1074	1.2706	0.0160	0.6348	2.0078	0.0143	0.0984
		4000	0.9970	0.0000	0.0130	0.9942	0.0000	0.1161	3.4944	0.0000	0.1068	3.6639	0.0833	0.1055	1.2326	0.0170	0.6298	2.0124	0.0143	0.0980
		8000	0.9985	0.0000	0.0043	0.9972	0.0000	0.0748	3.4940	0.0000	0.0918	3.6637	0.0833	0.1059	1.0383	0.0159	0.6176	2.0115	0.0143	0.0969
	3	2000	0.9898	0.0000	0.1423	0.9323	0.0000	0.2575	3.4904	0.0000	0.1321	3.6599	0.0769	0.1008	1.0752	0.0150	0.6311	2.0089	0.0121	0.0979
		4000	0.9954	0.0000	0.0759	0.9797	0.0000	0.2367	3.4897	0.0000	0.0998	3.6568	0.0769	0.1025	1.2851	0.0158	0.6266	2.0082	0.0121	0.0979
		8000	0.9977	0.0000	0.0196	0.9909	0.0000	0.0912	3.4932	0.0000	0.0828	3.6611	0.0769	0.0998	1.4331	0.0171	0.6202	2.0120	0.0121	0.0977
	4	2000	0.9881	0.0000	0.3168	0.7860	0.0000	0.3586	3.4892	0.0000	0.1111	3.6572	0.0714	0.0935	1.2057	0.0162	0.6304	2.0064	0.0105	0.0972
		4000	0.9942	0.0000	0.2656	0.8854	0.0000	0.3034	3.4904	0.0000	0.0884	3.6560	0.0714	0.0965	1.3948	0.0163	0.6264	2.0134	0.0105	0.0971
		8000	0.9972	0.0000	0.1584	0.9356	0.0000	0.2104	3.4941	0.0000	0.0784	3.6585	0.0714	0.0966	1.1489	0.0157	0.6171	2.0088	0.0105	0.0977
Scenario 2	2	2000	0.9972	0.0000	0.0002	0.9883	0.0000	0.0000	0.9985	0.0000	0.0000	2.7182	0.0833	0.0291	1.0647	0.0156	0.0820	2.6802	0.0143	0.0290
		4000	0.9988	0.0000	0.0000	0.9945	0.0000	0.0000	0.9994	0.0000	0.0000	2.7172	0.0833	0.0286	0.9938	0.0156	0.0813	2.6815	0.0143	0.0289
		8000	0.9992	0.0000	0.0000	0.9971	0.0000	0.0000	0.9995	0.0000	0.0000	2.7210	0.0833	0.0277	0.9967	0.0165	0.0802	2.6863	0.0143	0.0285
	3	2000	0.9953	0.0000	0.0008	0.9320	0.0000	0.0000	0.9980	0.0000	0.0000	2.6872	0.0769	0.0303	1.0855	0.0154	0.0814	2.6761	0.0121	0.0297
		4000	0.9980	0.0000	0.0000	0.9811	0.0000	0.0000	0.9992	0.0000	0.0000	2.6869	0.0769	0.0308	0.9935	0.0156	0.0811	2.6818	0.0121	0.0285
		8000	0.9989	0.0000	0.0000	0.9913	0.0000	0.0000	0.9995	0.0000	0.0000	2.7014	0.0769	0.0277	0.9967	0.0167	0.0803	2.6826	0.0121	0.0291
	4	2000	0.9939	0.0000	0.0012	0.7858	0.0000	0.0000	0.9990	0.0000	0.0000	2.6670	0.0714	0.0325	0.9879	0.0156	0.0810	2.6789	0.0105	0.0289
		4000	0.9971	0.0000	0.0000	0.8932	0.0000	0.0000	0.9995	0.0000	0.0000	2.6764	0.0714	0.0301	0.9939	0.0158	0.0811	2.6823	0.0105	0.0286
		8000	0.9986	0.0000	0.0000	0.9478	0.0000	0.0000	0.9996	0.0000	0.0000	2.6661	0.0714	0.0327	0.9968	0.0167	0.0803	2.6777	0.0105	0.0295

1. Varying-Coef represents for our variable selection method; GroupL represents for group LASSO method;
   LASSO1 represents for applying LASSO method to model (3.1); LASSO2 represents for applying LASSO method to model (3.2);
   SW1 represents for applying stepwise method to model (3.1); SW2 represents for applying stepwise method to model (3.2).
2. Note that the estimated RME should converge to the standard deviation of ε_i (i.e., 1 in our MC design). Therefore,
   an estimated RME closing to 1 is an indicator for good model performance of the corresponding method.

including lifestyle factors, such as physical activity (Galani and Schneider, 2007), alcohol consumption (Colditz et al., 1991), smoking habits (Cawley and Scholder, 2013) and so on; socio-economic factors (Cohen et al., 2013) such as education, income, working arrangement, etc.; and some other factors such as marital status (Sobal, Rauschenbach and Frongillo, 1992), duration of US residence (Oza-Frank and Cunningham, 2010), and depression (Faith et al., 2011). As discussed, a range of previous studies show that the impacts of regressors X on BMI are varying across demographic groups (Colditz et al., 1991; Sobal, Rauschenbach and Frongillo, 1992; Zhang and Wang, 2004). Therefore, we choose categorical variables of age, gender and ethnicity as covariates, i.e., Z in our model. By excluding such individuals with underweight and those having missing values of any variable involved in the model, we end up with a data set having 16593 observations. Definitions and summary statistics for all variables are presented in Table 2. Furthermore, Table 3 lists all 32 (i.e., $m = 32$) possible realizations of the covariates.

4.2. *Summary of the Main Findings.*

4.2.1. *Variable Selection.* First of all, we implement (2.5) to estimate the optimal bandwidth parameters. Results are reported in Table 4. It can be seen that all three covariates are relevant, however, their influences on the impacts of regressors on BMI are quite different. In particular, ethnicity and gender have relatively stronger influences than age group because the smoothing parameters associated with *ethnicity* and *sex* are much smaller than that of *age*.

Based on these smoothing parameters, we then apply our method to identify the relevant and irrelevant regressors to BMI. The optimal weight parameter selected by the modified BIC-type criterion through (2.13) is $\hat{\tilde{\gamma}} = 3.2$. Table 5 presents the result of variable selection through equation (2.7). 24 regressors, out of 48 in total, are identified as relevant, and the others are irrelevant to BMI.

In particular, while our estimate suggests that exercise is correlated with BMI, the level of intensity and frequency does matter. For example, compared to never doing vigorous (or strength) activity, doing such a level of exercise less than once per week has almost no effect on BMI, while doing it more than once per week starts to change BMI. In terms of light/moderate activity, however, people have to do it more than three times per week to see some effect on BMI. Results from our study may provide guidance for policy

available in NHIS.

TABLE 2. *Data Description and Summary Statistics*

Variable	Definition	Mean	St.D
Y			
BMI	body mass index	27.96	6.01
Z			
sex	0 for female and 1 for male	0.49	0.50
age	0 for age<25, 1 for 25<=age<=44, 2 for 45<=age<=64, and 3 for age>=65	1.39	0.75
race	0 for white, 1 for black, 2 for asian, 3 for all the other races	0.33	0.67
X			
Lifestyle factors			
vig_l0	1 if never do vigorous activities, 0 otherwise (reference group)	0.45	0.50
vig_l1	1 if do vigorous activities less than once per week, 0 otherwise	0.04	0.19
vig_l2	1 if do vigorous activities more than one time and less than three times per week, 0 otherwise	0.28	0.45
vig_l3	1 if do vigorous activities more than three times per week, 0 otherwise	0.23	0.42
mod_l0	1 if never do light/moderate activities, 0 otherwise (reference group)	0.35	0.48
mod_l1	1 if do light/moderate activities less than once per week, 0 otherwise	0.02	0.15
mod_l2	1 if do light/moderate activities more than one time and less than three times per week, 0 otherwise	0.29	0.46
mod_l3	1 if do light/moderate activities more than three times per week, 0 otherwise	0.33	0.47
str_l0	1 if never do strength activities, 0 otherwise (reference group)	0.66	0.47
str_l1	1 if do strength activities less than once per week, 0 otherwise	0.02	0.14
str_l2	1 if do strength activities more than one time and less than three times per week, 0 otherwise	0.20	0.40
str_l3	1 if do strength activities more than three times per week, 0 otherwise	0.12	0.32
smk_ed	1 if current every day smoker, 0 otherwise	0.13	0.34
smk_sd	1 if current some day smoker, 0 otherwise	0.04	0.20
smk_f	1 if former smoker, 0 otherwise	0.20	0.40
smk_n	1 if never smoke, 0 otherwise (reference group)	0.62	0.48
cigsday	number of cigarettes per day	1.98	5.52
alc1yr	1 if Ever had 12+ drinks in any one year, 0 otherwise	0.72	0.45
alc_life	1 if Had 12+ drinks in entire life, 0 otherwise	0.13	0.33
alc_c0	1 if do not drink at all currently, 0 otherwise (reference group)	0.26	0.44
alc_c1	1 if current infrequent drinker, 0 otherwise	0.12	0.33
alc_c2	1 if current light drinker, 0 otherwise	0.36	0.48
alc_c3	1 if current moderate drinker, 0 otherwise	0.19	0.39
alc_c4	1 if current heavier drinker, 0 otherwise	0.06	0.25
cpuse_0	1 if never or almost never use computer, 0 otherwise (reference group)	0.15	0.35
cpuse_1	1 if use computer for some/most days, 0 otherwise	0.18	0.38
cpuse_2	1 if use computer on every day, 0 otherwise	0.67	0.47
Socio-economic factors			
educ1	number of years of school completed	15.54	3.08
occup1	1 if management, business, science, and arts occupations, 0 otherwise	0.38	0.49
occup2	1 if service occupations, 0 otherwise	0.18	0.38
occup3	1 if sales and office occupations, 0 otherwise	0.23	0.42
occup4	1 if natural resources, construction, and maintenance occupations, 0 otherwise	0.09	0.29
occup5	1 if production, transportation, and material moving occupations, 0 otherwise (reference group)	0.12	0.33
working	1 if working or with job last week, 0 otherwise	0.88	0.32
unemp	1 if looking for job last week, 0 otherwise	0.05	0.21
nowork	1 if not working at a job last week, 0 otherwise	0.05	0.22
retired	1 if retired, 0 otherwise (reference group)	0.02	0.15
wrkhrs	hours worked last week	35.46	17.28
lnincome	nature logrithm of total earnings last year	10.20	0.94
houseown	1 if own or being bought the house, 0 otherwise	0.56	0.50
notcov	1 if not have health insurance coverage, 0 otherwise	0.20	0.40
hp	1 if ever seen/talked to health professional in the last 12 months, 0 otherwise	0.79	0.40
hce_l1	1 if amount family spent for medical care is 0, 0 otherwise (reference group)	0.13	0.33
hce_l2	1 if amount family spent for medical care is less than $500 but more than 0, 0 otherwise	0.37	0.48
hce_l3	1 if amount family spent for medical care is less than $1999 but more than $500, 0 otherwise	0.30	0.46
hce_l4	1 if amount family spent for medical care is less than $2999 but more than $2000, 0 otherwise	0.09	0.29
hce_l5	1 if amount family spent for medical care is less than $4999 but more than $3000, 0 otherwise	0.06	0.24
hce_l6	1 if amount family spent for medical care is $5000 or more, 0 otherwise	0.06	0.23
Other factors			
married	1 if married or de facto, 0 otherwise	0.51	0.50
us_born	1 if born in the US, 0 otherwise	0.81	0.39
us_m15	1 if stay in the US for more than 15 years, 0 otherwise	0.12	0.32
us_m5l15	1 if stay in the US for more than 5 years but less than 15 years, 0 otherwise	0.06	0.24
us_l5	1 if stay in the US for less than 5 years, 0 otherwise (reference group)	0.02	0.12
citizenp	1 if U.S. citizen, 0 otherwise	0.90	0.30
mental	1 if have depression/anxiety/emotional problem, 0 otherwise	0.01	0.12
rg_ne	1 if live in north east, 0 otherwise	0.16	0.37
rg_mw	1 if live in midwest, 0 otherwise	0.21	0.41
rg_sth	1 if live in south, 0 otherwise	0.36	0.48
rg_west	1 if live in west, 0 otherwise (reference group)	0.27	0.44

imsart-aoas ver. 2014/10/16 file: aoasmainfinal.tex date: March 3, 2017

TABLE 3. *List of realizations of covariates in the data and the percentage of observations for each group*

	Male										Female									
	Age				Ethnicity				Perc		Age				Ethnicity				Perc	
GI	<25	[25,45)	[45,65)	>=65	W	B	A	O		GI	<25	[25,45)	[45,65)	>=65	W	B	A	O		
1	x				x				3.9%	17	x				x				4.1%	
2	x					x			1.0%	18	x					x			0.7%	
3	x						x		0.3%	19	x						x		0.3%	
4	x							x	0.1%	20	x							x	0.1%	
5		x			x				17.0%	21		x			x				17.9%	
6		x				x			4.3%	22		x				x			2.9%	
7		x					x		1.6%	23		x					x		1.9%	
8		x						x	0.4%	24		x						x	0.4%	
9			x		x				14.6%	25			x		x				14.4%	
10			x			x			3.1%	26			x			x			2.3%	
11			x				x		1.0%	27			x				x		1.1%	
12			x					x	0.2%	28			x					x	0.3%	
13				x	x				2.6%	29				x	x				2.5%	
14				x		x			0.4%	30				x		x			0.2%	
15				x			x		0.1%	31				x			x		0.1%	
16				x				x	0.1%	32				x				x	0.1%	

GI = Group Index
Perc = Percentage of the whole sample
M = Male, F = Female
W = White, B = Black, A = Asian, O = Other

TABLE 4. *Estimated bandwidths for covariates*

sex	0.1158	*age group*	0.1979	*ethnicity*	0.0703

TABLE 5. *List of relevant and irrelevant variables to BMI*

Relevant variable	**Irrelevant variable**
lifestyle factors	*lifestyle factors*
vig_l2	vig_l1
vig_l3	mod_l1
mod_l3	mod_l2
str_l2	str_l1
str_l3	smk_sd
smk_ed	cpuse_1
smk_f	cpuse_2
cigsday	*socio-economic factors*
alc1yr	occup3
alc_life	occup4
alc_c1	working
alc_c2	unemp
alc_c3	nowork
alc_c4	wrkhrs
socio-economic factors	houseown
educ1	notcov
occup1	hce_l2
occup2	hce_l3
lnincome	hce_l4
hp	hce_l5
other factors	hce_l6
us_born	*other factors*
us_m15	us_m5l15
rg_sth	citizenp
married	rg_ne
mental	rg_mw

makers to adopt more efficient incentives to avoid overweight or obesity, i.e., encouraging people to do more intensive exercise or to do moderate exercise

more frequently rather than simply promoting exercise at any intensive level with any frequency.

Both the status of drinking and smoking and their consumption level are relevant to BMI. No impact from computer use can be seen. For socio-economic factors, education, income, and the two highest levels of occupational social class (OSC) (*occup1* and *occup2*, compared to lowest OSC, i.e., *occup5*), and health professional visit in the last 12 months are identified as relevant regressors for BMI, but the two lower levels of OSC (*occup3* and *occup4*, compared to *occup5*), working arrangement, working hours, house ownership, health insurance coverage and medical care expenditure are irrelevant to BMI. Among all other factors, indicators on duration of living in the U.S. (i.e., born in the U.S. and living in the U.S. more than 15 years, compared to living in the U.S. less than 5 years), living in the south (compared to living in the west), marital status and mental health problems are robust factors for BMI, however living in the US more than 5 years but less than 15 years (compared to less than 5 years), citizenship, living in either the north east or the middle west (compared to living in the west) have no impact on BMI.

For comparison purposes, in this BMI study we also estimate the other five models applied in Section 3, i.e., group LASSO method, LASSO method applied to models (3.1) and (3.2), respectively, and stepwise method applied to models (3.1) and (3.2), respectively. X and Z in models (3.1) and (3.2) have the same specification as what has been discussed in Section 4.1. It is worthwhile to mention that such variables selected by our method are exactly the same as those selected by group LASSO method. To compare model performance, we calculate root leave-one-out mean squared prediction errors (RME) $RME = (\sum_{i=1}^{N}(\hat{y}_{-i} - y_i)^2/N)^{1/2}$ for each model in Table 6[8], where $\hat{y}_{-i}$ denotes the leave-one-out prediction for the ith individual. It can be seen that our method outperforms all the other five models with the lowest RME. It is also interesting to see that group LASSO method performs as the second best, followed by LASSO methods applied to model (3.2) (the one taking account of varying impacts of X on BMI through interaction terms between X and Z). The LASSO method applied to model (3.1) (i.e., no varying impact is accounted for) performs worse than its counterpart. Performance of stepwise method is the worst amongst all options. Besides the superior performance of our method, these results also demonstrate, to some extent, that the varying impacts of potential factors on BMI are widely

[8]We also calculate RME for each of the 32 demographic groups from each method. Because of space limitation, these results are provided in the supplementary file (Gao et al., 2017).

TABLE 6. *Model Comparison on RME*

	Vary-Coeff	GroupL	LASSO1	SW1	LASSO2	SW2
RME	0.1562	0.1609	0.1657	0.2714	0.1646	0.2846

presented.

4.2.2. *Varying Impacts.* To quantify the effects of relevant regressors on BMI, we conduct a post-selection estimation using the unregularized estimation method for the varying-coefficient model only including the relevant regressors (i.e., equation (2.9)). For the sake of space limitation, in the supplementary file (Gao et al., 2017) we provide the full estimation results, including point and confidence interval estimates for the relevant determinants' impacts on BMI across demographic groups. Generally speaking, these estimated coefficients confirm that the selected variables are truly relevant to BMI. Because none of these regressors have their effects over all 32 groups to be constant zero, given zero is not consistently covered by the, at least 95%, CIs[9] of the 32 varying-effects of each regressor.

Taking the regressor of *us_born* as an example, its varying effects on BMI cross 32 demographic groups are shown in Figure 1. The demographic groups are indicated in the horizontal axis (for details, see Table 3). "×" represents the point estimate from the post-selection estimation, and the vertical line represents the 95% CI estimate. Two results emerge from this figure. First, the post-selection results show that the estimated effects of *us_born* on BMI are positive for all groups, which confirms that the regressor of *us_born* is truly relevant to BMI. Second, the effects of *us_born* on BMI are apparently varying across the 32 demographic groups. In particular, the effects are higher for males (groups 1-16) than females (groups 17-32) when age and race are the same, i.e., group 1 vs group 17, 2 vs 18, and so forth. Furthermore, the differences are more significant for Asian groups. As shown in Figure 1, there is almost no overlap between the two corresponding CI estimates, i.e., group 3 vs group 19, 7 vs 23, 11 vs 27, and 15 vs 31. Comparing across groups having the same gender and age range, *us_born* normally has higher impacts

[9]We cannot obtain CI's for the estimates provided in (2.7). After using the procedure of variable selection, following Wang and Xia (2009), we are able to calculate the 95% CIs through bootstrap for the post-selection estimates. See Theorem 2 and the discussions under Theorem 4 of Li, Ouyang and Racine (2013) for details. We point out that these CI's should be interpreted with caution. Indeed, these CI estimates might not be reliable without further justifying the variable selection bias issue. One sufficient condition for the validity of post-selection CIs is that all true relevant regressors are successfully identified by (2.7). We refer readers to Dezeure et al. (2015) and Bühlmann and Mandozzi (2014) for other sufficient conditions with further theoretical justification.

for Asian people. Taking the four youngest male groups as an example, being born in US increases BMI by 12.78% for Asians, which is higher than the increases of 6.11%, 11.24%, and 8.69% for white, black and all other races, respectively.

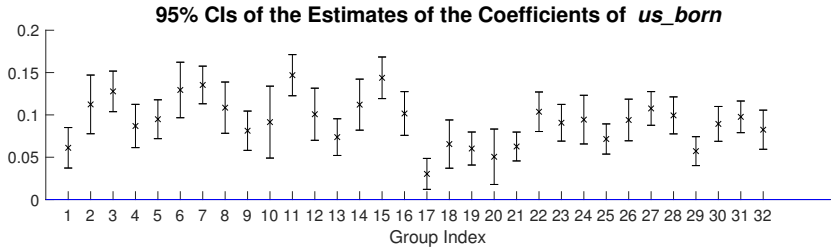

FIG 1. *The post-selection estimates for a relevant regressor of us_born*

5. Conclusions with Discussions. In order to solve some challenging modelling and statistical issues existing in the literature of BMI studies, we propose a variable selection procedure for the categorical varying-coefficient model. We examine the impacts of a wide range of potential factors proposed in the huge literature on BMI and obesity by using data from the 2013 NHIS in the United States. Specifically, (1) we allow for and quantify the varying impacts of determinants on BMI by using a varying-coefficient setting; (2) we systematically justify the relative importance of demographic variables in differencing potential determinants' impacts on BMI by looking at the optimal bandwidths of demographic group variables; (3) we identify the relatively important determinants of BMI by using a group LASSO technique.

Correspondingly, we also derive some asymptotic properties for the data-driven procedure documented in this paper. Our theoretical results show that true model can be successfully detected with probability going to 1 under certain mild conditions. In addition, the proposed estimator also achieves asymptotic normality on the true (oracle) model, whenever there is no irrelevant covariate.

In this study, we have not investigated any asymptotic behaviour for the case where both p and r diverge to infinity. If we ignore the optimal bandwidth selection by using the indicator function to replace all kernel functions and let p and r diverge to infinity (let alone the fact that the number of demographic groups grows exponentially with r), the theoretical study reduces to that investigated by Lounici et al. (2011). However, to the best knowledge of authors, how to achieve the optimal bandwidths for model (2.2) remains unknown for the high-dimensional case. We will pursue this

in a future study.

In the end, as suggested by one referee, it is worthwhile to mention that quantile regression model (Koenker, 2005) is an alternative approach if the interest is in some specific range (e.g., low or high) of BMI observations. In fact, a similar variable selection problem under the quantile categorical varying-coefficient model is considered by Zhao, Zhang and Liu (2014). Through using a penalized approach with both LASSO and fused LASSO (Tibshirani et al., 2005) penalties, their method particularly advocates the fusion of categories of determinants for each regressor, hence less emphasizing varying impacts among different categories, which is the focus of our approach via a group LASSO penalty. The major difference between the proposed quantile regression procedure in Zhao, Zhang and Liu (2014) and our method is that the former cannot justify the relative importance of demographic variables while our method achieves this goal by adopting a kernel function to select optimal bandwidth in (2.5). For studies particularly interested in specific ranges of BMI, it would be more interesting to enable the corresponding quantile categorical varying-coefficient model to retrieve the information of demographic variables by properly marrying a bandwidth selection procedure and group LASSO type penalty. We leave it as a future project.

Appendix A: Assumptions.

Assumption 1:

1. $\{X_i, Z_i, Y_i\}_{i=1}^N$ are i.i.d. In addition, $\max_{\bar{z}\in\bar{\mathcal{D}}} \|\beta_0(\bar{z})\| < \infty$.
2. $E[Y_i^2|X_i = x, \bar{Z}_i = \bar{z}]$ is bounded on $(x, \bar{z}) \in \mathbb{R}^p \times \bar{\mathcal{D}}$.
3. Let $\sigma_\varepsilon^2(x, \bar{z}) = E[\varepsilon_i^2|X_i = x, \bar{Z}_i = \bar{z}]$ and $\sigma_\varepsilon^2(\bar{z}) = E[\sigma_\varepsilon^2(X_i, \bar{z})|\bar{Z}_i = \bar{z}]$. Then $E[\sigma_\varepsilon^2(X_i, \bar{z})X_iX_i'|\bar{Z}_i = \bar{z}]$ is positive definite for all $\bar{z} \in \bar{\mathcal{D}}$.
4. For $s = 1, \ldots, r$, the sth component of $z = (z_1, \ldots, z_r)'$ takes c_s different values in $\{0, 1, \ldots, c_s - 1\}$. Moreover, $2 \le \min_{1\le s\le r} c_s \le \max_{1\le s\le r} c_s < \infty$.

Assumption 2:

1. **Relevant Covariate Case: i.e., $\bar{r} = r$**
   Define $L_{ij,\Theta} = L(Z_i, Z_j, \Theta)$, $m(Z_i) = E[X_iX_i'|Z_i]$ and

$$\eta_\beta(Z_j) = (E[X_iX_i'L_{ij,\Theta}|Z_j])^{-1}E[X_iX_i'\beta(Z_i)L_{ij,\Theta}|Z_j].$$

   Then $\Theta = 0_{r\times 1}$ are the only values of $\Theta = (\theta_1, \ldots, \theta_r)'$ that make

$$\sum_{z\in\mathcal{D}} \Pr(z)[\eta_\beta(z) - \beta_0(z)]'m(z)[\eta_\beta(z) - \beta_0(z)] = 0.$$

2. **Irrelevant Covariate Case: i.e., $\bar{r} < r$**
For $\tilde{Z}_i = (Z_{i,\bar{r}+1}, \ldots, Z_{i,r})'$, $\{\tilde{Z}_i, 1 \le i \le N\}$ is independent of all other variables and has no impact on $\beta_0(\cdot)$. Define $\bar{L}_{ij,\bar{\Theta}}$, $\bar{\eta}_\beta(\bar{Z}_j) = (E[X_i X_i' \bar{L}_{ij,\bar{\Theta}} | \bar{Z}_j])^{-1} E[X_i X_i' \beta(\bar{Z}_i) \bar{L}_{ij,\bar{\Theta}} | \bar{Z}_j]$ and $\bar{m}(\bar{Z}_i) = E[X_i X_i' | \bar{Z}_i]$. Then the only values of $\Theta = (\theta_1, \ldots, \theta_{\bar{r}})'$ that make

$$\sum_{\bar{z} \in \bar{\mathcal{D}}} \Pr(\bar{z}) [\bar{\eta}_\beta(\bar{z}) - \beta_0(\bar{z})]' m(\bar{z}) [\bar{\eta}_\beta(\bar{z}) - \beta_0(\bar{z})] = 0$$

are $\bar{\Theta} = 0_{\bar{r} \times 1}$. $\theta_s \in [0, 1]$ for $s = \bar{r} + 1, \ldots, r$.

Assumption 3:

1. For a random variable $\bar{Z}_i \in \bar{\mathcal{D}}$ and $\beta_0(\bar{Z}_i) = (\beta_{01}(\bar{Z}_i), \ldots, \beta_{0p}(\bar{Z}_i))'$, suppose there exists an integer $0 < p^* \le p$ such that $0 < E|\beta_{0j}(\bar{Z}_i)|^2 < \infty$ for $j = 1, \ldots, p^*$ and $E|\beta_{0j}(\bar{Z}_i)|^2 = 0$ for $j = p^* + 1, \ldots, p$.
2. For any $\bar{z} \in \bar{\mathcal{D}}$, $0 < \alpha_1 \le \rho_{\min} \le \rho_{\max} \le \alpha_2 < \infty$, where $\rho_{\min}$ and $\rho_{\max}$ denote the minimum and maximum eigenvalues of $E[X_i X_i' | \bar{z}]$ respectively, and α_1, α_2 are two universal positive constants.

Assumptions 1 and 2 are identical to those in Li, Ouyang and Racine (2013). Note that since the support $\mathcal{D}$ is finite, we automatically have $\Pr(z) = \Pr(Z_i = z) > \alpha_3 > 0$ with some universal constant α_3 for any $z \in \mathcal{D}$. Assumption 3.2 ensures all eigenvalues of $E[X_i X_i' | \bar{z}]$ are bounded uniformly.

SUPPLEMENTARY MATERIAL

Supplement to "Variable Selection for a Categorical Varying-Coefficient Model with Identifications for Determinants of Body Mass Index"
(doi: COMPLETED BY THE TYPESETTER). In this supplementary file, we provide a detailed presentation and discussion on (1) mathematical proofs of the main results, (2) estimation procedure of our method, (3) extra simulation results, and (4) other estimation results from the BMI study.

References.

Aitchison, J. and Aitken, C. (1976). Multivariate binary discrimination by the kernel method. *Biometrika* **63** 413–420.

Ali, S. M. and Lindström, M. (2006). Socioeconomic, psychosocial, behavioural, and psychological determinants of BMI among young women: Differing patterns for underweight and overweight/obesity. *The European Journal of Public Health* **16** 324–330.

Bühlmann, P. and Mandozzi, J. (2014). High-dimensional variable screening and bias in subsequent inference, with an empirical comparison. *Computational Statistics* **29** 407–430.

Carey, M., Small, H., Yoong, S. L., Boyes, A., Bisquera, A. and Sanson-Fisher, R. (2014). Prevalence of comorbid depression and obesity in general practice: A cross-sectional survey. *British Journal of General Practice* **64** e122–e127.

Cawley, J. (2011). *The Oxford Handbook of the Social Science of Obesity*. Oxford University Press.

Cawley, J. and Scholder, S. v. H. K. (2013). The demand for cigarettes as derived from the demand for weight control Technical Report, National Bureau of Economic Research.

Chu, W., Li, R. and Reimherr, M. (2016). Feature screening for time-varying coefficient models with ultrahigh dimensional longitudinal data. *The Annals of Applied Statistics* **10** 596–617.

Cohen, A. K., Rai, M., Rehkopf, D. H. and Abrams, B. (2013). Educational attainment and obesity: A systematic review. *Obesity Reviews* **14** 989–1005.

Colditz, G. A., Giovannucci, E., Rimm, E. B., Stampfer, M. J., Rosner, B., Speizer, F. E., Gordis, E. and Willett, W. C. (1991). Alcohol intake in relation to diet and obesity in women and men. *American Journal of Clinical Nutrition* **54** 49–55.

Dezeure, R., Bühlmann, P., Meier, L. and Meinshausen, N. (2015). High-Dimensional Inference: Confidence Intervals, p-Values and R-Software hdi. *Statistical Science* **30** 533–558.

Faith, M., Butryn, M., Wadden, T., Fabricatore, A., Nguyen, A. and Heymsfield, S. (2011). Evidence for prospective associations among depression and obesity in population-based studies. *Obesity Reviews* **12** e438–e453.

Fan, J. and Zhang, W. (1999). Statistical estimation in varying coefficient models. *The Annals of Statistics* **27** 1491–1518.

Fontaine, K. R., Redden, D. T., Wang, C., Westfall, A. O. and Allison, D. B. (2003). Years of life lost due to obesity. *JAMA* **289** 187–193.

Galani, C. and Schneider, H. (2007). Prevention and treatment of obesity with lifestyle interventions: Review and meta-analysis. *International Journal of Public Health* **52** 348–359.

Gao, J., Peng, B., Ren, Z. and Zhang, X. (2017). Supplement to "Variable selection for a categorical varying-coefficient model with identifications for determinants of body mass index".

Gertheiss, J. and Tutz, G. (2010). Sparse modeling of categorical explanatory variables. *The Annals of Applied Statistics* **4** 2150-2180.

Hall, P., Li, Q. and Racine, J. S. (2007). Nonparametric estimation of regression functions in the presence of irrelevant regressors. *Review of Economics and Statistics* **89** 784-789.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society. Series B* **55** 757–796.

Hill, J. O. and Peters, J. C. (1998). Environmental contributions to the obesity epidemic. *Science* **280** 1371–1374.

Huang, J., Ma, S., Xie, H. and Zhang, C.-H. (2009). A group bridge approach for variable selection. *Biometrika* **96** 339–355.

Koenker, R. (2005). *Quantile regression* **38**. Cambridge University Press.

Li, Q., Ouyang, D. and Racine, J. S. (2013). Categorical semiparametrics varying–coefficient models. *Journal of Applied Econometrics* **28** 551–579.

Li, Q. and Racine, J. S. (2010). Smooth varying–coefficient estimation and inference for qualitative and quantitative data. *Econometric Theory* **26** 1607–1637.

Lipowicz, A., Gronkiewicz, S. and Malina, R. M. (2002). Body mass index, overweight and obesity in married and never married men and women in Poland. *American Journal*

of Human Biology **14** 468–475.

Lounici, K., Pontil, M., van de Geer, S. and Tsybakov, A. B. (2011). Oracle inequalities and optimal inference under group sparsity. *The Annals of Statistics* **39** 2164–2204.

Ma, S., Carroll, R. J., Liang, H. and Xu, S. (2015). Estimation and inference in generalized additive coefficient models for nonlinear interactions with high-dimensional covariates. *The Annals of Statistics* **43** 2102-2131.

Oza-Frank, R. and Cunningham, S. A. (2010). The weight of US residence among immigrants: A systematic review. *Obesity Reviews* **11** 271–280.

Rehkopf, D. H., Laraia, B. A., Segal, M., Braithwaite, D. and Epel, l. (2011). The relative importance of predictors of body mass index change, overweight and obesity in adolescent girls. *International Journal of Pediatric Obesity* **6** 233–242.

Sobal, J., Rauschenbach, B. S. and Frongillo, E. A. (1992). Marital status, fatness and obesity. *Social Science & Medicine* **35** 915–923.

Stice, E., Shaw, H. and Marti, C. N. (2006). A meta-analytic review of obesity prevention programs for children and adolescents: The skinny on interventions that work. *Psychological Bulletin* **132** 667.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67** 91–108.

Von Kries, R., Toschke, A. M., Koletzko, B. and Slikker, W. (2002). Maternal smoking during pregnancy and childhood obesity. *American Journal of Epidemiology* **156** 954–961.

Wang, H. and Leng, C. (2007). Unified LASSO estimation by least squares approximation. *Journal of the American Statistical Association* **102** 1039–1048.

Wang, L., Li, H. and Huang, J. Z. (2008). Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *Journal of the American Statistical Association* **103** 1556–1569.

Wang, H. and Xia, Y. (2009). Shrinkage estimation of the varying coefficient. *Journal of the American Statistical Association* **104** 747—757.

WHO (2015). Obesity and overweight Fact Sheet No. 311., Working paper available at http://www.who.int/mediacentre/factsheets/fs311/en/.

Yu, Y. (2012). Educational differences in obesity in the United States: A closer look at the trends. *Obesity* **20** 904–908.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B* **68** 49–67.

Zeng, W., Eisenberg, D. T., Jovel, K. R., Undurraga, E. A., Nyberg, C., Tanner, S., Reyes-García, V., Leonard, W. R., Castano, J., Huanca, T. et al. (2013). Adult obesity: Panel study from native Amazonians. *Economics & Human Biology* **11** 227–235.

Zhang, Q. and Wang, Y. (2004). Socioeconomic inequality of obesity in the United States: Do gender, age, and ethnicity matter? *Social Science & Medicine* **58** 1171–1180.

Zhao, W., Zhang, R. and Liu, J. (2014). Regularization and model selection for quantile varying coefficient model with categorical effect modifiers. *Computational Statistics & Data Analysis* **79** 44–62.

Department of Econometrics
and Business Statistics
Monash University
VIC 3145, Australia
E-mail: Jiti.Gao@monash.edu

Department of Economics
University of Bath
Bath BA2 7JP, UK
E-mail: bp495@bath.ac.uk.

Department of Statistics
University of Pittsburgh
Pittsburgh, PA 15260.
E-mail: zren@pitt.edu

Department of Economics
University of Exeter
Exeter EX4 4PU, UK
E-mail: x.zhang1@exeter.ac.uk

Submitted to the Annals of Applied Statistics

SUPPLEMENT TO "VARIABLE SELECTION FOR A CATEGORICAL VARYING-COEFFICIENT MODEL WITH IDENTIFICATIONS FOR DETERMINANTS OF BODY MASS INDEX"

By Jiti Gao*, Bin Peng, Zhao Ren and Xiaohui Zhang

Monash University, University of Bath, University of Pittsburgh and University of Exeter

Appendix S1: Mathematical Proofs. This appendix contains the proofs of the theorems presented in the paper and the below Lemma S1.1. Throughout this appendix, diag(V) denotes a square diagonal matrix with the elements of the vector V on the main diagonal; $\to_P$ denotes converging in probability; $\to_D$ denotes converging in distribution; $\|\cdot\|$ denotes the Euclidean norm; let $\Pr(z) = \Pr(Z_i = z)$ for notational simplicity when no misunderstanding can arise; i_p denotes a $p \times 1$ one vector; and I_p denotes a $p \times p$ identity matrix.

For notational simplicity, partition $\hat{\Theta}$ as $\hat{\Theta} = (\hat{\bar{\Theta}}', \hat{\tilde{\Theta}}')'$ conformably with $Z_i = (\bar{Z}_i', \tilde{Z}_i')'$, where $\hat{\bar{\Theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_{\bar{r}})'$ and $\hat{\tilde{\Theta}} = (\hat{\theta}_{\bar{r}+1}, \ldots, \hat{\theta}_r)'$. In addition, for $j = 1, \ldots, m$, let $\bar{z}^j = (z_1^j, \ldots, z_{\bar{r}}^j)'$ and $\tilde{z}^j = (z_{\bar{r}+1}^j, \ldots, z_r^j)'$. In the following proofs, we will repeatedly use these symbols.

Lemma S1.1. *Let $\hat{\Theta} = (\hat{\bar{\Theta}}', \hat{\tilde{\Theta}}')'$ be the one obtained from Lemma 2.1. The notion $\Delta_1(\tilde{z})$ is denoted in (S1.2) below. Under Assumptions 1-3, for $\forall z \in \mathcal{D}$, we have*

1. $\frac{1}{N}\sum_{i=1}^N X_i X_i' L(Z_i, z, \hat{\Theta}) = \Pr(\bar{z}) E[X_i X_i' | \bar{z}] \Delta_1(\tilde{z}) + O_P\left(\frac{1}{\sqrt{N}}\right)$;
2. $\frac{1}{N}\sum_{i=1}^N X_i X_i' \beta(\bar{Z}_i) L(Z_i, z, \hat{\Theta}) = \Pr(\bar{z}) E[X_i X_i' \beta(\bar{z}) | \bar{z}] \Delta_1(\tilde{z}) + O_P\left(\frac{1}{\sqrt{N}}\right)$;
3. $\frac{1}{N}\sum_{i=1}^N X_i \varepsilon_i L(Z_i, z, \hat{\Theta}) = O_P\left(\frac{1}{\sqrt{N}}\right)$;
4. $\frac{1}{N}\sum_{i=1}^N X_i' \beta_0(Z_i) \varepsilon_i L(Z_i, z, \hat{\Theta}) = O_P\left(\frac{1}{\sqrt{N}}\right)$;
5. $\frac{1}{N}\sum_{i=1}^N \varepsilon_i^2 L(Z_i, z, \hat{\Theta}) = \sigma_\varepsilon^2(\bar{z}) \Pr(\bar{z}) \Delta_1(\tilde{z}) + O_P\left(\frac{1}{\sqrt{N}}\right)$.

Proof of Lemma S1.1:

1): By the definition of the kernel function used in this paper, we can write for $s = 1, \ldots, r$

*Jiti Gao acknowledges the Australian Research Council Discovery Grants Program support under Grant numbers: DP150101012 & DP170104421.

$$l(Z_{i,s}, z_s, \hat{\theta}_s) = 1(Z_{i,s} = z_s) + \hat{\theta}_s 1(Z_{i,s} \neq z_s).$$

Based on Lemma 2.1, we can simplify the product kernel as

$$L(Z_i, z, \hat{\Theta}) = \left(1(\bar{Z}_i = \bar{z}) + \sum_{s=1}^{\bar{r}} \hat{\theta}_s 1_{s,\bar{Z}_i=\bar{z}} + O_P(\|\hat{\bar{\Theta}}\|^2) \right) L(\tilde{Z}_i, \tilde{z}, \hat{\tilde{\Theta}}),$$

where $1_{s,\bar{Z}_i=\bar{z}} = 1(Z_{i,s} \neq z_s) \prod_{n=1,n\neq s}^{\bar{r}} 1(Z_{i,n} = z_n)$. Therefore,

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N} X_i X_i' L(Z_i, z, \hat{\Theta}) \\
&= \frac{1}{N}\sum_{i=1}^{N} X_i X_i' \left(1(\bar{Z}_i = \bar{z}) + \sum_{s=1}^{\bar{r}} \hat{\theta}_s 1_{s,\bar{Z}_i=\bar{z}} + O_P(\|\hat{\bar{\Theta}}\|^2) \right) \tilde{L}(\tilde{Z}_i, \tilde{z}, \hat{\tilde{\Theta}}) \\
&\equiv A_1 + A_2 + A_3, \qquad \text{(S1.1)}
\end{aligned}$$

where

$$\begin{aligned}
A_1 &= \frac{1}{N}\sum_{i=1}^{N} X_i X_i' 1(\bar{Z}_i = \bar{z}) \tilde{L}(\tilde{Z}_i, \tilde{z}, \hat{\tilde{\Theta}}), \\
A_2 &= \frac{1}{N}\sum_{i=1}^{N} X_i X_i' \sum_{s=1}^{\bar{r}} \hat{\theta}_s 1_{s,\bar{Z}_i=\bar{z}} \tilde{L}(\tilde{Z}_i, \tilde{z}, \hat{\tilde{\Theta}}), \\
A_3 &= O_P(\|\hat{\bar{\Theta}}\|^2) \frac{1}{N}\sum_{i=1}^{N} X_i X_i' \tilde{L}(\tilde{Z}_i, \tilde{z}, \hat{\tilde{\Theta}}).
\end{aligned}$$

Notice that we can expand the product form of $\tilde{L}(\tilde{Z}_i, \tilde{z}, \hat{\tilde{\Theta}})$ as a summation form:

$$\begin{aligned}
\tilde{L}(\tilde{Z}_i, \tilde{z}, \hat{\tilde{\Theta}}) &= \prod_{s=\bar{r}+1}^{r} (1(Z_{i,s} = z_s) + \hat{\theta}_s 1(Z_{i,s} \neq z_s)) \\
&= \prod_{s=\bar{r}+1}^{r} 1(Z_{i,s} = z_s) + \cdots + \prod_{s=\bar{r}+1}^{r} \hat{\theta}_s 1(Z_{i,s} \neq z_s).
\end{aligned}$$

Then, for simplicity, we denote

$$\Delta_1(\tilde{z}) = E\left[\prod_{s=\bar{r}+1}^{r} 1(Z_{i,s} = z_s) \right] + \cdots + \prod_{s=\bar{r}+1}^{r} \hat{\theta}_s E\left[\prod_{s=\bar{r}+1}^{r} 1(Z_{i,s} \neq z_s) \right] \qquad \text{(S1.2)}$$

as the expectation of $\tilde{L}(\tilde{Z}_i, \tilde{z}, \hat{\tilde{\Theta}})$ with respect to $\tilde{Z}_i$. In connection with Assumption 1, it is easy to know that

$$A_1 = \Pr(\bar{z}) E[X_i X_i' | \bar{z}] \Delta_1(\tilde{z}) + O_P\left(\frac{1}{\sqrt{N}} \right),$$

where $\Delta_1(\tilde{z})$ is denoted in (S1.2).

For A_2,

$$\begin{aligned}\|A_2\| &\leq \sum_{s=1}^{\bar{r}} \left\| \theta_s \frac{1}{N} \sum_{i=1}^{N} X_i X_i' 1_{s,\bar{Z}_i=\bar{z}} \tilde{L}(\tilde{Z}_i, \tilde{z}, \hat{\tilde{\Theta}}) \right\| \\ &\leq \sum_{s=1}^{\bar{r}} |\hat{\theta}_s| \left\| \frac{1}{N} \sum_{i=1}^{N} X_i X_i' 1_{s,\bar{Z}_i=\bar{z}} \tilde{L}(\tilde{Z}_i, \tilde{z}, \hat{\tilde{\Theta}}) \right\| \\ &\leq O_P(\|\hat{\tilde{\Theta}}\|) \sum_{s=1}^{\bar{r}} \left\| \frac{1}{N} \sum_{i=1}^{N} X_i X_i' 1_{s,\bar{Z}_i=\bar{z}} \tilde{L}(\tilde{Z}_i, \tilde{z}, \hat{\tilde{\Theta}}) \right\| = O_P(\|\hat{\tilde{\Theta}}\|).\end{aligned}$$

Similar to A_2, we can show that $A_3 = O_P(\|\hat{\tilde{\Theta}}\|^2)$. Based on Lemma 2.1 and the analysis for A_1, A_2 and A_3, the proof is completed. ■

2)-5): The results follow from the procedure similar to 1) of this lemma. ■

Proof of Theorem 2.1:

1): Let $\alpha_N = \frac{1}{\sqrt{N}}$ and U be an $m \times p$ matrix. We want to show that for any given $\epsilon > 0$, there exists a large constant C such that

$$\liminf_N \Pr\left\{ \inf_{\|U\|=C} Q_\gamma\left(B_0 + \alpha_N U\right) > Q_\gamma(B_0) \right\} \geq 1 - \epsilon. \tag{S1.3}$$

This implies with probability at least $1-\epsilon$ that there exists a local minimum in the ball $\{B_0 + \alpha_N U : \|U\| \leq C\}$. Hence, there exists a local minimizer such that $\|\hat{B} - B_0\| = O_P(\alpha_N)$. The above argument is in line with the same spirit as the proofs for Theorem 1 of Fan and Li (2001) and Lemma A.1 of Wang and Xia (2009).

For notational simplicity, let U_j be the transpose of the jth row of the matrix U with $j = 1, \ldots, m$ and V_s be the sth column of the matrix U with $s = 1, \ldots, p$; and denote

$$\begin{aligned}e_j &= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} X_i \left(X_i' \beta_0(\bar{Z}_i) - X_i' \beta_0(\bar{z}^j) + \varepsilon_i \right) L(Z_i, z^j, \hat{\Theta}) \\ &= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} X_i X_i' \left(\beta_0(\bar{Z}_i) - \beta_0(\bar{z}^j) \right) L(Z_i, z^j, \hat{\Theta}) + \frac{1}{\sqrt{N}} \sum_{i=1}^{N} X_i \varepsilon_i L(Z_i, z^j, \hat{\Theta}).\end{aligned}$$

By result (3) of Lemma S1.1, it is easy to know that $\frac{1}{\sqrt{N}} \sum_{i=1}^{N} X_i \varepsilon_i L(Z_i, z^j, \hat{\Theta}) = O_P(1)$ uniformly in j. We now focus on the next term:

$$\begin{aligned}&\left\| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} X_i X_i' \left(\beta_0(\bar{Z}_i) - \beta_0(\bar{z}^j) \right) L(Z_i, z^j, \hat{\Theta}) \right\| \\ =& \left\| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} X_i X_i' \left(\beta_0(\bar{Z}_i) - \beta_0(\bar{z}^j) \right) \right.\end{aligned}$$

$$
\begin{aligned}
&\times \left(1(\bar{Z}_i = \bar{z}^j) + \sum_{s=1}^{\bar{r}} \hat{\theta}_s 1_{s,\bar{Z}_i=\bar{z}^j} + O_P(\|\hat{\tilde{\Theta}}\|^2) \right) \tilde{L}(\tilde{Z}_i, \tilde{z}, \hat{\tilde{\Theta}}) \Bigg\| \\
&\leq \left\| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} X_i X_i' \left(\beta_0(\bar{Z}_i) - \beta_0(\bar{z}^j) \right) \sum_{s=1}^{\bar{r}} \hat{\theta}_s 1_{s,\bar{Z}_i=\bar{z}^j} \tilde{L}(\tilde{Z}_i, \tilde{z}, \hat{\tilde{\Theta}}) \right\| \\
&\quad + \left\| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} X_i X_i' \left(\beta_0(\bar{Z}_i) - \beta_0(\bar{z}^j) \right) O_P(\|\hat{\tilde{\Theta}}\|^2) \tilde{L}(\tilde{Z}_i, \tilde{z}, \hat{\tilde{\Theta}}) \right\| = O_P(1),
\end{aligned}
\tag{S1.4}
$$

where the last equality follows from Lemma 2.1, Assumption 1, and results (1) and (2) of Lemma S1.1. Therefore, we know that $e_j = O_P(1)$ uniformly in j due to the fact that $\mathcal{D}$ is compact.

Then we write

$$
\begin{aligned}
&Q_\gamma(B_0 + \alpha_N U) - Q_\gamma(B_0) \\
&= \sum_{j=1}^{m} \sum_{i=1}^{N} \left(X_i'\beta_0(\bar{Z}_i) + \varepsilon_i - X_i'\beta_0(\bar{z}^j) - \alpha_N X_i' U_j \right)^2 L(Z_i, z^j, \hat{\Theta}) \\
&\quad + \sum_{s=1}^{p^*} \gamma_s \|b_{0s} + \alpha_N V_s\| + \sum_{s=p^*+1}^{p} \gamma_s \|\alpha_N V_s\| \\
&\quad - \sum_{j=1}^{m} \sum_{i=1}^{N} \left(X_i'\beta_0(\bar{Z}_i) + \varepsilon_i - X_i'\beta_0(\bar{z}^j) \right)^2 L(Z_i, z^j, \hat{\Theta}) - \sum_{s=1}^{p^*} \gamma_s \|b_{0s}\| \\
&= \sum_{j=1}^{m} \sum_{i=1}^{N} \left(\alpha_N X_i' U_j \right)^2 L(Z_i, z^j, \hat{\Theta}) + \sum_{s=1}^{p^*} \gamma_s \left(\|b_{0s} + \alpha_N V_s\| - \|b_{0s}\| \right) \\
&\quad + \sum_{s=p^*+1}^{p} \gamma_s \|\alpha_N V_s\| - 2 \sum_{j=1}^{m} \sum_{i=1}^{N} \alpha_N U_j' X_i \left(X_i'\beta_0(\bar{Z}_i) - X_i'\beta_0(\bar{z}^j) + \varepsilon_i \right) L(Z_i, z^j, \hat{\Theta}) \\
&\geq \sum_{j=1}^{m} \sum_{i=1}^{N} \alpha_N^2 U_j' X_i X_i' U_j L(Z_i, z^j, \hat{\Theta}) + \sum_{s=1}^{p^*} \gamma_s \left(\|b_{0s} + \alpha_N V_s\| - \|b_{0s}\| \right) \\
&\quad - 2 \sum_{j=1}^{m} \sum_{i=1}^{N} \alpha_N U_j' X_i \left(X_i'\beta_0(\bar{Z}_i) - X_i'\beta_0(\bar{z}^j) + \varepsilon_i \right) L(Z_i, z^j, \hat{\Theta}) \\
&\geq \sum_{j=1}^{m} \|U_j\|^2 \frac{\Lambda_{min}(z^j)}{2} - 2 \sum_{j=1}^{m} U_j' e_j + \sum_{s=1}^{p^*} \gamma_s \left(\|b_{0s} + \alpha_N V_s\| - \|b_{0s}\| \right) \\
&\geq \sum_{j=1}^{m} \|U_j\|^2 \frac{\Lambda_{min}(z^j)}{2} - 2 \sum_{j=1}^{m} U_j' e_j - O(1) \sum_{s=1}^{p^*} \gamma_s \frac{1}{\sqrt{N}} \|V_s\|,
\end{aligned}
$$

where $\Lambda_{min}(z^j)$ denotes the minimum eigenvalue of $\Pr(\bar{z}^j) E[X_i X_i' | \bar{z}^j] \Delta_1(\tilde{z}^j)$; the

second inequality follows from (1) of Lemma S1.1; the third inequality follows from the Mean Value Theorem. Notice that $\|U\| = C$, so we further write that

$$
\begin{aligned}
& Q_\gamma(B_0 + \alpha_N U) - Q_\gamma(B_0) \\
\geq & \sum_{j=1}^{m} \|U_j\|^2 \frac{\Lambda_{min}(z^j)}{2} - 2\sum_{j=1}^{m} U_j' e_j - O(1)\sum_{s=1}^{p^*} \gamma_s \frac{1}{\sqrt{N}} \|V_s\| \\
\geq & \sum_{j=1}^{m} \|U_j\|^2 \frac{\Lambda_{min}(z^j)}{2} - 2\left(\sum_{j=1}^{m} \|U_j\|^2 \sum_{j=1}^{m} \|e_j\|^2\right)^{1/2} - O(1)\sum_{s=1}^{p^*} \gamma_s \frac{1}{\sqrt{N}} \|V_s\| \\
\geq & C^2 \min_j \frac{\Lambda_{min}(z^j)}{2} - 2C\left(\sum_{j=1}^{m} \|e_j\|^2\right)^{1/2} - O(1)\frac{1}{\sqrt{N}}\|\gamma^*\| \left(\sum_{s=1}^{p^*} \|V_s\|^2\right)^{1/2} \\
= & C^2 \min_j \frac{\Lambda_{min}(z^j)}{2} - 2C\left(\sum_{j=1}^{m} \|e_j\|^2\right)^{1/2} - O(1)C,
\end{aligned}
\tag{S1.5}
$$

where we have used that $\frac{1}{\sqrt{N}}\|\gamma^*\| = O(1)$ by the condition in the body of this theorem and $\|e_j\| = O_P(1)$ uniformly in j. Notice that $\frac{1}{2}C^2 \min_j \Lambda_{min}(z^j)$ is a quadratic function of C while the rest terms on RHS of (S1.5) are linear in C. Since C can be sufficiently large, it is easy to know that RHS of (S1.5) is positive with probability arbitrarily close to 1. The proof for (S1.3) is now completed. ■

2): For simplicity, we show that $\Pr(\|\hat{b}_{\gamma,p}\| = 0) \to 1$ only. The proofs for $\hat{b}_{\gamma,j}$ with $j = p^* + 1, \ldots, p-1$ are the same. If $\|\hat{b}_{\gamma,p}\| \neq 0$, $\hat{B}$ must satisfy the following equation:

$$
0 = \frac{\partial}{\partial b_p} Q_\gamma(B) = A_1 + A_2, \tag{S1.6}
$$

where

$$
A_1 = -\sum_{i=1}^{N} 2X_{i,p}\left((Y_i - X_i'\hat{\beta}_{\gamma,1})L(Z_i, z^1, \hat{\Theta}), \ldots, (Y_i - X_i'\hat{\beta}_{\gamma,m})L(Z_i, z^m, \hat{\Theta})\right)',
$$

and $A_2 = \frac{\gamma_p}{\|b_p\|} b_p$. For $s = 1, \ldots, m$, we can further write each element of A_1 as follows:

$$
\begin{aligned}
\frac{A_{1,s}}{\sqrt{N}} &= \frac{-1}{\sqrt{N}} \sum_{i=1}^{N} 2X_{i,p}\left(X_i'(\beta_0(\bar{Z}_i) - \hat{\beta}_{\gamma,s}) + \varepsilon_i\right) L(Z_i, z^s, \hat{\Theta}) \\
&= \frac{-1}{\sqrt{N}} \sum_{i=1}^{N} 2X_{i,p}X_i'(\beta_0(\bar{Z}_i) - \hat{\beta}_{\gamma,s}) L(Z_i, z^s, \hat{\Theta}) - \frac{1}{\sqrt{N}} \sum_{i=1}^{N} 2X_{i,p}\varepsilon_i L(Z_i, z^s, \hat{\Theta})
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{-1}{\sqrt{N}} \sum_{i=1}^{N} 2X_{i,p} X_i' (\beta_0(\bar{Z}_i) - \beta_0(z^s)) L(Z_i, z^s, \hat{\Theta}) \\
&\quad - \frac{1}{\sqrt{N}} \sum_{i=1}^{N} 2X_{i,p} X_i' (\beta_0(z^s) - \hat{\beta}_{\gamma,s}) L(Z_i, z^s, \hat{\Theta}) + O_P(1) \\
&= O_P(1),
\end{aligned}
$$

where the third equality follows from (3) of Lemma S1.1; the last equality follows from (S1.4) and the first result of this theorem.

On the other hand, $\left\| \frac{1}{\sqrt{N}} A_2 \right\| = \frac{1}{\sqrt{N}} \gamma_p \geq \frac{1}{\sqrt{N}} \min_{s \in \{p^*+1,\ldots,p\}} \gamma_s \geq \omega_2$ by the condition in the body of this theorem, where ω_2 is sufficiently large. Therefore, $\Pr(\|A_1\| < \|A_2\|) \to 1$, which implies that, with probability tending to 1, (S1.6) does not hold. The above analysis implies that $\hat{b}_{\gamma,p}$ must be located at the place where the objective function (2.5) is not differentiable with respect to b_p. Since equation (2.5) of the main file is only not differentiable with respect to b_p at the origin, we immediately obtain that $\Pr(\|\hat{b}_{\gamma,p}\| = 0) \to 1$. The same procedure of proof applies to $\hat{b}_{\gamma,j}$ with $j = p^* + 1, \ldots, p-1$. The proof is then completed. ■

Proof of Theorem 2.2:

By Theorem 2.1, we know that $\hat{\beta}_{\gamma,js} = 0$ for $j = 1, \ldots, m$ and $s = p^* + 1, \ldots, p$ with probability tending to one. After using some simple algebra, we can obtain the first derivative of equation (2.7) of the main file with respect to β_j for $j = 1, \ldots, m$. Then it is easy to know that $\hat{\beta}_{\gamma,jU}$ must be the solution of the following equation:

$$
\frac{2}{N} \sum_{i=1}^{N} X_{iU} \left(Y_i - X_{iU}' \hat{\beta}_{\gamma,jU} \right) L(Z_i, z^j, \hat{\Theta}) - \frac{1}{N} D \hat{\beta}_{\gamma,jU} = 0,
$$

where $\hat{\beta}_{\gamma,jU} = (\hat{\beta}_{\gamma,j1}, \ldots, \hat{\beta}_{\gamma,jp^*})'$ and $D = \operatorname{diag}\left(\gamma_1 \|\hat{b}_{\gamma,1}\|^{-1}, \ldots, \gamma_{p^*} \|\hat{b}_{\gamma,p^*}\|^{-1} \right)$. It implies that $\hat{\beta}_{\gamma,jU}$ must have the form:

$$
\hat{\beta}_{\gamma,jU} = \left(\frac{1}{N} \sum_{i=1}^{N} X_{iU} X_{iU}' L(Z_i, z^j, \hat{\Theta}) + \frac{1}{2N} D \right)^{-1} \frac{1}{N} \sum_{i=1}^{N} X_{iU} Y_i L(Z_i, z^j, \hat{\Theta}).
$$

Comparing with the oracle estimator,

$$
\text{(S1.7)} \qquad \left\| \hat{\beta}_{\gamma,jU} - \hat{\beta}_{ora}(\bar{z}^j) \right\| \leq \left\| \Sigma_N(z^j) \right\| \left\| \frac{1}{N} \sum_{i=1}^{N} X_{iU} Y_i L(Z_i, z^j, \hat{\Theta}) \right\|,
$$

where $\Sigma_N(z^j)$ is denoted as

$$
\begin{aligned}
&\Sigma_N(z^j) \\
&= \left(\frac{1}{N} \sum_{i=1}^{N} X_{iU} X_{iU}' L(Z_i, z^j, \hat{\Theta}) + \frac{1}{2N} D \right)^{-1} - \left(\frac{1}{N} \sum_{i=1}^{N} X_{iU} X_{iU}' L(Z_i, z^j, \hat{\Theta}) \right)^{-1}.
\end{aligned}
$$

Since $\Sigma_N(z^j)$ has a finite dimension, it is easy to know that the rate of $\left\|\Sigma_N(z^j)\right\|$ converging to 0 is the same as

$$\left\| \frac{1}{N}\sum_{i=1}^{N} X_{iU}X_{iU}'L(Z_i, z^j, \hat{\Theta}) + \frac{1}{2N}D - \frac{1}{N}\sum_{i=1}^{N} X_{iU}X_{iU}'L(Z_i, z^j, \hat{\Theta}) \right\|$$
$$= \left\| \frac{1}{2N}D \right\| = O_P\left(\frac{\|\gamma^*\|}{N}\right).$$

Moreover, by (2) and (4) of Lemma S1.1, $\frac{1}{N}\sum_{i=1}^{N} X_{iU}Y_iL(Z_i, z^j, \hat{\Theta}) = O_P(1)$. Therefore, for $j = 1, \ldots, m$, $\left\|\hat{\beta}_{\gamma,jU} - \hat{\beta}_{ora}(\bar{z}^j)\right\| = O_P\left(\frac{\|\gamma^*\|}{N}\right)$. We therefore complete the proof. ■

Proof of Theorem 2.3:

(1): For an arbitrary model S, we say it is under-fitted if it misses at least one variable with a nonzero coefficient (i.e., $S \subset U$ but $U \neq S$); it is over fitted if S covers all relevant variables but also includes at least one redundant regressor (i.e., $U \subset S$ but $U \neq S$). Then, according to whether the model S_γ is under fitted, correctly fitted, or over fitted, we create three mutually exclusive sets $A^- = \{\tilde{\gamma} \in \mathbb{R} : S_{\tilde{\gamma}} \subset U, S_{\tilde{\gamma}} \neq U\}$, $A^0 = \{\tilde{\gamma} \in \mathbb{R} : S_{\tilde{\gamma}} = U\}$ and $A^+ = \{\tilde{\gamma} \in \mathbb{R} : S_{\tilde{\gamma}} \supset U, S_{\tilde{\gamma}} \neq U\}$. Suppose that $\tilde{\beta}_j$ for $j = 1, \ldots, m$ are the unpenalized estimators and there is a sequence $\{\hat{\gamma}_N\}$ that ensures (2.10) of the main file satisfies the conditions required by Theorem 2.1. For example, say those used in the section of Monte Carlo study.

Case 1: Under-fitted model, i.e., $S \subset U$ but $U \neq S$. Without loss of generality, we assume that only one variable is missing, so we assume that the first $p^* - 1$ elements of $\hat{\beta}_{\tilde{\gamma},j}$ are obtained from the under–fitted model and the rest $p - p^* + 1$ elements of $\hat{\beta}_{\tilde{\gamma},j}$ are 0.

We then write

$$\begin{aligned}
RSS_{\tilde{\gamma}} &= \frac{1}{N}\sum_{j=1}^{m}\sum_{i=1}^{N}\left(Y_i - X_i'\hat{\beta}_{\tilde{\gamma},j}\right)^2 L(Z_i, z^j, \hat{\Theta}) \\
&= \frac{1}{N}\sum_{j=1}^{m}\sum_{i=1}^{N}\left(Y_i - X_i'\tilde{\beta}_j + X_i'\tilde{\beta}_j - X_i'\hat{\beta}_{\tilde{\gamma},j}\right)^2 L(Z_i, z^j, \hat{\Theta}) \\
&= \frac{1}{N}\sum_{j=1}^{m}\sum_{i=1}^{N}\left(Y_i - X_i'\tilde{\beta}_j\right)^2 L(Z_i, z^j, \hat{\Theta}) \\
&\quad + \frac{1}{N}\sum_{j=1}^{m}\sum_{i=1}^{N}\left(X_i'\tilde{\beta}_j - X_i'\hat{\beta}_{\tilde{\gamma},j}\right)^2 L(Z_i, z^j, \hat{\Theta}) \\
&\quad + \frac{2}{N}\sum_{j=1}^{m}\sum_{i=1}^{N}\left(\tilde{\beta}_j - \hat{\beta}_{\tilde{\gamma},j}\right)' X_i\left(Y_i - X_i'\tilde{\beta}_j\right) L(Z_i, z^j, \hat{\Theta})
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{N}\sum_{j=1}^{m}\sum_{i=1}^{N}\left(Y_i - X_i'\tilde{\beta}_j\right)^2 L(Z_i, z^j, \hat{\Theta}) \\
&\quad + \frac{1}{N}\sum_{j=1}^{m}\sum_{i=1}^{N}\left(X_i'\tilde{\beta}_j - X_i'\hat{\beta}_{\tilde{\gamma},j}\right)^2 L(Z_i, z^j, \hat{\Theta}) \\
&\equiv RSS^* + R_{2\tilde{\gamma}}
\end{aligned}
$$

where the fourth equality is due to

$$
\frac{2}{N}\sum_{j=1}^{m}\sum_{i=1}^{N}\left(\tilde{\beta}_j - \hat{\beta}_{\tilde{\gamma},j}\right)' X_i\left(Y_i - X_i'\tilde{\beta}_j\right) L(Z_i, z^j, \hat{\Theta}) = 0
$$

by the definition of unpenalized estimators.

We now consider $R_{2\tilde{\gamma}}$ and write

$$
\begin{aligned}
R_{2\tilde{\gamma}} &= \frac{1}{N}\sum_{j=1}^{m}\sum_{i=1}^{N}\left(\tilde{\beta}_j - \hat{\beta}_{\tilde{\gamma},j}\right)' X_i X_i' L(Z_i, z^j, \hat{\Theta})\left(\tilde{\beta}_j - \hat{\beta}_{\tilde{\gamma},j}\right) \\
&= \sum_{j=1}^{m}\left(\tilde{\beta}_j - \hat{\beta}_{\tilde{\gamma},j}\right)' \Sigma_1(z^j)\left(\tilde{\beta}_j - \hat{\beta}_{\tilde{\gamma},j}\right) + O_P\left(\frac{1}{\sqrt{N}}\right) \\
&\geq \sum_{j=1}^{m}\Lambda_{min}(z^j)\left\|\tilde{\beta}_j - \hat{\beta}_{\tilde{\gamma},j}\right\|^2 + O_P\left(\frac{1}{\sqrt{N}}\right) \\
&= O(1)\sum_{j=1}^{m}\left\|\tilde{\beta}_j - \hat{\beta}_{\tilde{\gamma},j}\right\|^2 + O_P\left(\frac{1}{\sqrt{N}}\right) \\
&\geq O(1)\sum_{j=1}^{m}\tilde{\beta}_{j,p^*}^2 + O_P\left(\frac{1}{\sqrt{N}}\right) \gg O_P\left(\frac{1}{\sqrt{N}}\right),
\end{aligned}
$$

where $\Sigma_1(z^j) = \Pr(\bar{z}^j)E[X_iX_i'|\bar{z}^j]\Delta_1(\tilde{z}^j)$ and $\Delta_1(\tilde{z}^j)$ is denoted in (S1.2); $\Lambda_{min}(z^j)$ denotes the minimum eigenvalue of $\Pr(\bar{z}^j)E[X_iX_i'|\bar{z}^j]\Delta_1(\tilde{z}^j)$; $\tilde{\beta}_{j,p^*}$ denotes the p^*th element of $\tilde{\beta}_j$; the second equality follows from (1) of Lemma S1.1 of the Appendix; the first inequality follows from Assumption 3.

Similarly, we can obtain that $RSS_{\hat{\gamma}_N} \equiv RSS^* + R_{2\hat{\gamma}_N}$, where

$$
\begin{aligned}
R_{2\hat{\gamma}_N} &= \frac{1}{N}\sum_{j=1}^{m}\sum_{i=1}^{N}\left(\tilde{\beta}_j - \hat{\beta}_{\hat{\gamma}_N,j}\right)' X_i X_i' L(Z_i, z^j, \hat{\Theta})\left(\tilde{\beta}_j - \hat{\beta}_{\hat{\gamma}_N,j}\right) \\
&= \sum_{j=1}^{m}\left(\tilde{\beta}_j - \hat{\beta}_{\hat{\gamma}_N,j}\right)' \Sigma_1(z^j)\left(\tilde{\beta}_j - \hat{\beta}_{\hat{\gamma}_N,j}\right) + O_P\left(\frac{1}{\sqrt{N}}\right) \\
&\leq \sum_{j=1}^{m}\Lambda_{max}(z^j)\left\|\tilde{\beta}_j - \hat{\beta}_{\hat{\gamma}_N,j}\right\|^2 + O_P\left(\frac{1}{\sqrt{N}}\right)
\end{aligned}
$$

$$
\begin{aligned}
&\leq O\left(\sum_{j=1}^{m}\left\|\tilde{\beta}_j-\hat{\beta}_{\hat{\gamma}_N,j}\right\|^2\right)+O_P\left(\frac{1}{\sqrt{N}}\right)\\
&\leq O\left(\sum_{j=1}^{m}\left\|\tilde{\beta}_j-\beta_0(\bar{z}^j)\right\|^2+\sum_{j=1}^{m}\left\|\hat{\beta}_{\hat{\gamma}_N,j}-\beta_0(\bar{z}^j)\right\|^2\right)+O_P\left(\frac{1}{\sqrt{N}}\right)\\
&=O_P\left(\frac{1}{\sqrt{N}}\right),
\end{aligned}
$$

where $\Lambda_{max}(z^j)$ denotes the maximum eigenvalue of $\Pr(\bar{z}^j)E[X_iX_i'|\bar{z}^j]\Delta_1(\tilde{z}^j)$; the second equality follows from (1) of Lemma S1.1 of the Appendix; the first inequality follows from Assumption 3; the last equality follows from using Theorem 2.1 on both $\sum_{j=1}^{m}\left\|\tilde{\beta}_j-\beta_0(\bar{z}^j)\right\|^2$ and $\sum_{j=1}^{m}\left\|\hat{\beta}_{\hat{\gamma}_N,j}-\beta_0(\bar{z}^j)\right\|^2$.

Note by (5) of Lemma S1.1 we can obtain $RSS^*\to_P\sum_{j=1}^{m}\sigma_\varepsilon^2(\bar{z}^j)\Pr(\bar{z}^j)\Delta_1(\tilde{z}^j)$. Based on the analysis on $R_{2\tilde{\gamma}}$ and $R_{2\hat{\gamma}_N}$, we then can further conclude that

$$
\Pr\left(\inf_{\tilde{\gamma}\in A^-}BIC_{\tilde{\gamma}}>BIC_{\hat{\gamma}_N}\right)\to 1.
$$

Case 2: Over-fitted model, i.e., $S\supset U$ but $U\neq S$. Consider $\forall\tilde{\gamma}\in A^+$ and recall that $\hat{B}_{\tilde{\gamma}}$ determines a model $S_{\tilde{\gamma}}$. Under such a model $S_{\tilde{\gamma}}$, we can define another unpenalized estimate $\check{B}_{\tilde{\gamma}}$ as

$$
\check{B}_{\tilde{\gamma}}=\underset{\beta_1,\ldots,\beta_m}{\operatorname{argmin}}\frac{1}{N}\sum_{j=1}^{m}\sum_{i=1}^{N}\left(Y_i-X_i'\beta_j\right)^2L(Z_i,z^j,\hat{\Theta}),
$$

where, for $j=1,\ldots,m$, $\|\beta_{j,s}\|=0$ with $\forall s\notin S_{\tilde{\gamma}}$. In other words, $\check{B}_{\tilde{\gamma}}=(\check{\beta}_1,\ldots,\check{\beta}_m)'$ is the unpenalized estimator under the model determined by $\hat{B}_{\tilde{\gamma}}$. By definition, we obtain immediately that $RRS_{\tilde{\gamma}}\geq RRS_{S_{\tilde{\gamma}}}$, where

$$
RRS_{S_{\tilde{\gamma}}}=\frac{1}{N}\sum_{j=1}^{m}\sum_{i=1}^{N}\left(Y_i-X_i'\check{\beta}_j\right)^2L(Z_i,z^j,\hat{\Theta}).
$$

It follows that

$$
\begin{aligned}
&\ln RRS_{\tilde{\gamma}}-\ln RRS^*\geq\ln RRS_{S_{\tilde{\gamma}}}-\ln RRS^*\\
&=\ln\left\{\frac{RRS^*}{RRS^*}+\frac{1}{N\cdot RRS^*}\sum_{j=1}^{m}\sum_{i=1}^{N}\left(\tilde{\beta}_j-\check{\beta}_j\right)'X_iX_i'L(Z_i,z^j,\hat{\Theta})\left(\tilde{\beta}_j-\check{\beta}_j\right)\right\}\\
&\geq-\frac{O(1)}{N\cdot RRS^*}\sum_{j=1}^{m}\sum_{i=1}^{N}\left(\tilde{\beta}_j-\check{\beta}_j\right)'X_iX_i'L(Z_i,z^j,\hat{\Theta})\left(\tilde{\beta}_j-\check{\beta}_j\right)\\
&\geq-\frac{O_P(1)}{RRS^*}\sum_{j=1}^{m}\Lambda_{max}(z^j)\left\|\tilde{\beta}_j-\check{\beta}_j\right\|^2
\end{aligned}
$$

$$\geq -\frac{O_P(1)}{RRS^*}\sum_{j=1}^{m}\Lambda_{max}(z^j)\left\|\tilde{\beta}_j-\beta_0(\bar{z}^j)\right\|^2-\frac{O_P(1)}{RRS^*}\sum_{j=1}^{m}\Lambda_{max}(z^j)\left\|\beta_0(\bar{z}^j)-\check{\beta}_j\right\|^2$$
$$\geq -\left|O_P\left(\frac{1}{N}\right)\right|,$$

where $\Lambda_{max}(z^j)$ denotes the maximum eigenvalue of $\Pr(\bar{z}^j)E[X_iX_i'|\bar{z}^j]\Delta_1(\bar{z}^j)$; $\tilde{\beta}_j$ for $j=1,\ldots,m$ are the same unpenalized estimators as those used in **Case 1**; the second inequality follows from (1) of Lemma S1.1 and Assumption 3; the fourth inequality follows from an application of result (1) in Theorem 2.1 on both terms $\sum_{j=1}^{m}\Lambda_{max}(z^j)\left\|\tilde{\beta}_j-\beta_0(\bar{z}^j)\right\|^2$ and $\sum_{j=1}^{m}\Lambda_{max}(z^j)\left\|\beta_0(\bar{z}^j)-\check{\beta}_j\right\|^2$.

Similarly, we can obtain that $\ln RRS_{\hat{\gamma}_N}-\ln RRS^*=O_P\left(\frac{1}{N}\right)$. Thus, we obtain that

$$\ln RRS_{\tilde{\gamma}}-\ln RRS_{\hat{\gamma}_N}\geq -\left|O_P\left(\frac{1}{N}\right)\right|.$$

We then write

$$\inf_{\tilde{\gamma}\in A^+}BIC_{\tilde{\gamma}}-BIC_{\hat{\gamma}_N}=\ln RRS_{\tilde{\gamma}}-\ln RRS_{\hat{\gamma}_N}+(df_{\tilde{\gamma}}-df_{\hat{\gamma}_N})\frac{\ln N}{N}.$$

By Theorem 2.1, we know that $\Pr(df_{\hat{\gamma}_N}\to p^*)=1$. Since $\tilde{\gamma}\in A^+$, we must have that $\Pr(df_{\tilde{\gamma}}\geq p^*+1)\to 1$. Then it is clear

$$\Pr\left(\inf_{\tilde{\gamma}\in A^+}BIC_{\tilde{\gamma}}>BIC_{\hat{\gamma}_N}\right)\to 1.$$

Combining Cases 1 and 2, we obtain that $\Pr\left(\inf_{\tilde{\gamma}\in A^-\cup A^+}BIC_{\tilde{\gamma}}>BIC_{\hat{\gamma}_N}\right)\to 1$. It further indicates that $\Pr\left(S_{\hat{\tilde{\gamma}}}\to U\right)=1$. It then completes the proof.

(2-3): The proofs of the second and third results of this theorem follow by noticing that setting $\tilde{\gamma}$ to a large constant satisfies all the conditions required by Theorem 2.2 and the first result of this theorem. Thus, we have

$$\hat{\beta}_{\hat{\tilde{\gamma}},jU}-\beta_0(\bar{z}^j)=\hat{\beta}_{ora}(\bar{z}^j)-\beta_0(\bar{z}^j)+O_P\left(\frac{1}{N}\right).$$

Then the results follow from Theorems 2 and 4 of Li, Ouyang and Racine (2013). ■

Appendix S2: Estimation Procedure. The estimation procedure for implementing our method is described below.

Steps:

1. Minimize the cross-validation criterion function (2.5) in order to choose $\hat{\Theta}$.

2. Select $\tilde{\gamma}$ defined in (2.10) from a sufficient large set, say $[1,\ 2\sqrt{N}]$ by using grid search. For each choice of $\tilde{\gamma}$, implement the estimation proposed by (2.7) in a similar procedure proposed in Wang and Xia (2009). Define

$$\hat{B}_{\tilde{\gamma}}^{(n)} = (\hat{\beta}_{\tilde{\gamma},1}^{(n)}, \ldots, \hat{\beta}_{\tilde{\gamma},m}^{(n)})' = (\hat{b}_{\tilde{\gamma},1}^{(n)}, \ldots, \hat{b}_{\tilde{\gamma},p}^{(n)}) \tag{S2.1}$$

to be the estimate obtained in the nth iteration. Then the loss function given above can be locally approximated by

$$\begin{aligned} &\sum_{j=1}^{m}\sum_{i=1}^{N}\left(Y_i - X_i'\beta_j\right)^2 L(Z_i, z^j, \hat{\Theta}) + \sum_{s=1}^{p}\tilde{\gamma}_s\frac{\|b_s\|^2}{\|\hat{b}_{\tilde{\gamma},s}^{(n)}\|} \\ &= \sum_{j=1}^{m}\left(\sum_{i=1}^{N}\left(Y_i - X_i'\beta_j\right)^2 L(Z_i, z^j, \hat{\Theta}) + \sum_{s=1}^{p}\tilde{\gamma}_s\frac{\beta_{j,s}^2}{\|\hat{b}_{\tilde{\gamma},s}^{(n)}\|}\right). \end{aligned} \tag{S2.2}$$

The minimizer of (S2.2) is provided by $\hat{B}_{\tilde{\gamma}}^{(n+1)} = (\hat{\beta}_{\tilde{\gamma},1}^{(n+1)}, \ldots, \hat{\beta}_{\tilde{\gamma},m}^{(n+1)})'$, where for $j = 1, \ldots, m$

$$\hat{\beta}_{\tilde{\gamma},j}^{(n+1)} = \left(\sum_{i=1}^{N} X_i X_i' L(Z_i, z^j, \hat{\Theta}) + D^{(n)}\right)^{-1}\sum_{i=1}^{N} X_i Y_i L(Z_i, z^j, \hat{\Theta}), \tag{S2.3}$$

and $D^{(n)} = \text{diag}\left(\|\hat{b}_{\tilde{\gamma},1}^{(n)}\|^{-1}\tilde{\gamma}, \ldots, \|\hat{b}_{\tilde{\gamma},p}^{(n)}\|^{-1}\tilde{\gamma}\right)$. Repeat this procedure until $\|\hat{B}_{\tilde{\gamma}}^{(n+1)} - \hat{B}_{\tilde{\gamma}}^{(n)}\| < tolerance$, where $tolerance$ is a sufficiently small number (say, 10^{-6}).

3. Select the optimal $\hat{\tilde{\gamma}}$ based on the modified BIC-type criterion (i.e., (2.13) of the main file of this paper).

4. After removing the irrelevant covariates and regressors, carry on the unregularized estimation as proposed in Li, Ouyang and Racine (2013).

Steps of Bootstrap Procedure of Post Selection Estimate for Empirical Study:

After we successfully remove the irrelevant regressors from X_i, we then implement the bootstrap steps documented in the simulation section of Li, Ouyang and Racine (2013). The detailed steps are as follows:

1. Select the optimal bandwidths using the relevant regressors and covariates.

2. After obtaining the optimal bandwidths, estimate the coefficients with unregularized estimators as documented in Li, Ouyang and Racine (2013).

3. Calculate the predicted value of the dependent variable $\widehat{Y}_i$ and the prediction error $\widehat{\varepsilon}_i = \widehat{Y}_i - Y_i$.

4. Start bootstrap replications. For each replication, we generate new dependent variables as $Y_i^* = \widehat{Y}_i + \widehat{\varepsilon}_i u_i$, where $u_i \sim N(0,1)$ and u_i is i.i.d. across i. For each replication, we implement Steps 1-2.

Notice that we also have tried using the same bandwidth obtained from Step 1 for all replications of Step 4. Compared to the results obtained by using optimal bandwidth in each replication, the results are almost identical for the first 3 decimals in this particular study. In order to save time, one can avoid implementing bandwidth selection in Step 4.

Appendix S3: Extra Discussion and Simulation Results. Model (2.2) nests the linear model, including interaction terms between X_i and Z_i as a special case. For example, let Z_i represent race having three categories, e.g., White, Black and Asian. X_i is a $p \times 1$ vector including constant as the first element. To generate interaction term between X_i and Z_i, we create two dummy variables as

$$Z_{1i} = \begin{cases} 1 & \text{if } i^{th} \text{ individual is White,} \\ 0 & \text{otherwise,} \end{cases},$$
$$Z_{2i} = \begin{cases} 1 & \text{if } i^{th} \text{ individual is Black,} \\ 0 & \text{otherwise.} \end{cases}.$$

One conventional way to capture the varying impacts of X on Y across Z is to apply the following specification:

$$Y_i = \mathcal{X}_i' \beta_0^* + \varepsilon_i, \tag{S3.1}$$

where $\mathcal{X}_i = (X_i',\ Z_{1i}X_i',\ Z_{2i}X_i')'$, and correspondingly $\beta_0^* = (\beta_{01}',\ \beta_{02}',\ \beta_{03}')'$. Then it can be rewritten as

$$\begin{aligned} Y_i &= X_i'\beta_{01} + Z_{1i}X_i'\beta_{02} + Z_{2i}X_i'\beta_{03} + \varepsilon_i \\ &= X_i'(\beta_{01} + Z_{1i}\beta_{02} + Z_{2i}\beta_{03}) + \varepsilon_i, \end{aligned} \tag{S3.2}$$

Note the second line of (S3.2) is a special case of (2.2) (by letting $\beta_0(Z_i) = \beta_{01} + Z_{1i}\beta_{02} + Z_{2i}\beta_{03}$), which allows for more complicated interactions, when we have an $r \times 1$ vector Z_i with $r \geq 1$.

To supplement results presented in the Monte Carlo study of the paper, we report the average computational time (seconds) for implementing the proposed method once in Matlab 2015b. Table S3.1 provides computational times on an Atlas cluster consists of 8 nodes with the following specifications:

2x 3.1GHz Intel Xeon E5-2687w v3 (10 Cores) 25MB L3 Cache 9.6GT/s QPI (Max Turbo Freq. 3.5GHz, Min 3.2GHz)

64GB 2133MHz ECC DDR4-RAM (Quad Channel)

2x 900GB 10,000 RPM SAS II Hard Drives (Raid) and 2x 1.2TB 10,000 RPM SAS II Hard Drives (Raid)

NIVDIA Quadro K2200 4GB Graphics Card (GPU)

Table S3.1: Average Computational Time for The Method Proposed Based on The Simulation Study

	N	Scenario 1.1	Scenario 1.2	Scenario 2
$r = 2$	2000	43.2262	56.5401	41.5559
	4000	191.2504	181.7986	122.9326
	8000	499.0367	574.1499	391.2775
$r = 3$	2000	99.4158	102.8270	67.7226
	4000	340.0857	399.1947	291.0900
	8000	1204.9186	2197.0378	1033.0460
$r = 4$	2000	125.3648	134.5891	97.7771
	4000	542.9951	809.5176	323.9053
	8000	2252.8226	2847.8641	2126.9989

For the purpose of demonstration, we plot the histograms of the estimates of bandwidth on irrelevant covariate (when $r = 2$) based on 1000 replications. As shown in Figure 1 below, the probability of removing the irrelevant covariate is around 55%, but it does not converge to 1 as the sample size goes up, which perfectly fits the second result of Lemma 2.1.

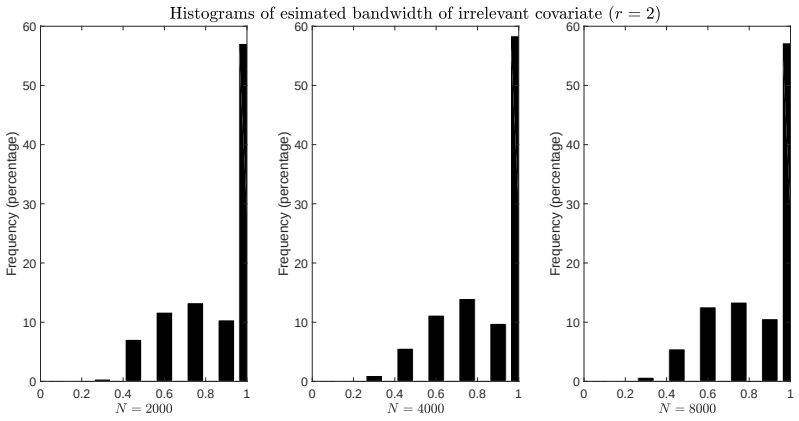

Fig 1: Histograms of estimated bandwidth of irrelevant covairate when $r = 2$

Appendix S4: Extra Results from BMI Study. In the paper, we present RME from each of the six methods and it can be seen that our method is superior to all the other competing methods in term of prediction accuracy. We report RME from all estimation methods over all 32 demographic groups in Table S4.1 to further demonstrate that our method outperform the others. Moreover, in Table S4.2 and S4.3, we present the full results from the post-selection estimation.

Table S4.1: RME over All Demographic Groups

Group	Vary-Coeff	GroupL	LASSO1	SW1	LASSO2	SW2
1	0.1854	0.1900	0.2028	0.2975	0.2012	0.2974
2	0.1612	0.1621	0.1784	0.2637	0.1787	0.2841
3	0.2168	0.2158	0.2526	0.3140	0.2516	0.3052
4	0.1313	0.1352	0.1563	0.2751	0.1543	0.3250
5	0.1675	0.1737	0.1781	0.2758	0.1769	0.2811
6	0.1443	0.1565	0.1514	0.2601	0.1505	0.2824
7	0.2130	0.2136	0.2282	0.3128	0.2254	0.3028
8	0.1702	0.1826	0.1761	0.2332	0.1791	0.2423
9	0.1579	0.1628	0.1658	0.2795	0.1647	0.2952
10	0.1457	0.1521	0.1502	0.2799	0.1500	0.3137
11	0.1958	0.2011	0.2067	0.2895	0.2069	0.2791
12	0.1500	0.1668	0.1491	0.3172	0.1480	0.3287
13	0.1525	0.1559	0.1614	0.2883	0.1609	0.3073
14	0.1288	0.1335	0.1462	0.2253	0.1432	0.2421
15	0.1807	0.1813	0.2009	0.2806	0.1955	0.2646
16	0.1781	0.1907	0.1575	0.3817	0.1636	0.4122
17	0.1690	0.1731	0.1782	0.2744	0.1769	0.2573
18	0.1820	0.1832	0.1887	0.2790	0.1850	0.2722
19	0.1693	0.1699	0.1931	0.2404	0.1889	0.2446
20	0.1296	0.1350	0.1384	0.2852	0.1426	0.2654
21	0.1374	0.1428	0.1413	0.2578	0.1403	0.2683
22	0.1410	0.1476	0.1489	0.2470	0.1472	0.2630
23	0.1575	0.1596	0.1615	0.2916	0.1595	0.2942
24	0.1441	0.1523	0.1589	0.2834	0.1629	0.2936
25	0.1317	0.1361	0.1351	0.2627	0.1341	0.2894
26	0.1378	0.1421	0.1454	0.2675	0.1453	0.3108
27	0.1362	0.1410	0.1421	0.2743	0.1395	0.2844
28	0.1407	0.1449	0.1380	0.2389	0.1385	0.2732
29	0.1316	0.1384	0.1335	0.2671	0.1352	0.2827
30	0.1149	0.1128	0.1232	0.2801	0.1215	0.3095
31	0.1576	0.1590	0.1728	0.2715	0.1794	0.3025
32	0.1376	0.1372	0.1404	0.0884	0.1207	0.1328

Appendix S4.2: Full Results from Post-Selection Estimation on Relevant Lifestyle Factors

Group Index	vig_l2	vig_l3	mod_l3	str_l2	str_l3	smk_ed	smk_f	cigsday	alc1yr	alc_life	alc_c1	alc_c2	alc_c3	alc_c4
1	-0.0171	-0.0416	-0.0181	-0.0323	-0.0276	0.0043	0.0239	-0.0001	0.0224	0.0395	0.0035	-0.0319	-0.0552	-0.0596
	(0.0077)[a]	(0.0087)	(0.0069)	(0.0085)	(0.0111)	(0.0218)	(0.0075)	(0.0013)	(0.0141)	(0.0133)	(0.0117)	(0.0120)	(0.0145)	(0.0151)
2	0.0123	0.0000	0.0034	-0.0402	-0.0780	0.0033	0.0241	-0.0013	0.0410	0.0785	-0.0455	-0.0282	-0.0466	-0.0642
	(0.0157)	(0.0169)	(0.0126)	(0.0161)	(0.0187)	(0.0271)	(0.0136)	(0.0018)	(0.0278)	(0.0198)	(0.0224)	(0.0248)	(0.0275)	(0.0302)
3	-0.0169	-0.0199	-0.0111	-0.0188	-0.0149	0.0363	0.0336	-0.0008	0.0421	0.0245	-0.0101	-0.0418	-0.0680	-0.1011
	(0.0107)	(0.0114)	(0.0101)	(0.0109)	(0.0127)	(0.0183)	(0.0113)	(0.0013)	(0.0177)	(0.0166)	(0.0157)	(0.0162)	(0.0189)	(0.0240)
4	-0.0056	-0.0195	-0.0156	-0.0432	-0.0362	0.0137	0.0199	-0.0006	0.0445	0.0426	-0.0133	-0.0443	-0.0717	-0.0773
	(0.0087)	(0.0113)	(0.0078)	(0.0099)	(0.0208)	(0.0182)	(0.0095)	(0.0011)	(0.0153)	(0.0143)	(0.0147)	(0.0136)	(0.0148)	(0.0169)
5	-0.0171	-0.0293	-0.0273	-0.0403	-0.0415	-0.0110	0.0202	0.0001	0.0123	0.0244	0.0054	-0.0323	-0.0652	-0.0743
	(0.0072)	(0.0085)	(0.0063)	(0.0079)	(0.0105)	(0.0173)	(0.0077)	(0.0011)	(0.0152)	(0.0134)	(0.0128)	(0.0125)	(0.0144)	(0.0154)
6	0.0181	-0.0247	-0.0101	-0.0286	-0.0365	0.0050	0.0121	-0.0024	0.0075	0.0367	-0.0086	-0.0163	-0.0409	-0.0279
	(0.0152)	(0.0151)	(0.0124)	(0.0159)	(0.0166)	(0.0258)	(0.0141)	(0.0018)	(0.0235)	(0.0219)	(0.0223)	(0.0221)	(0.0246)	(0.0302)
7	-0.0227	-0.0128	0.0025	-0.0106	-0.0246	0.0387	0.0241	-0.0018	0.0298	0.0061	0.0094	-0.0337	-0.0547	-0.0760
	(0.0104)	(0.0120)	(0.0092)	(0.0095)	(0.0162)	(0.0191)	(0.0108)	(0.0016)	(0.0198)	(0.0182)	(0.0174)	(0.0176)	(0.0205)	(0.0240)
8	-0.0127	-0.0229	-0.0174	-0.0405	-0.0333	0.0125	0.0187	-0.0010	0.0276	0.0155	-0.0070	-0.0461	-0.0807	-0.0757
	(0.0110)	(0.0103)	(0.0098)	(0.0118)	(0.0172)	(0.0202)	(0.0104)	(0.0013)	(0.0238)	(0.0181)	(0.0205)	(0.0230)	(0.0219)	(0.0266)
9	-0.0238	-0.0394	-0.0241	-0.0460	-0.0561	-0.0088	0.0118	-0.0014	0.0187	0.0124	0.0164	-0.0303	-0.0582	-0.0680
	(0.0073)	(0.0083)	(0.0061)	(0.0078)	(0.0102)	(0.0168)	(0.0064)	(0.0010)	(0.0138)	(0.0120)	(0.0124)	(0.0125)	(0.0136)	(0.0151)
10	-0.0250	-0.0209	-0.0040	-0.0322	-0.0881	-0.0077	0.0094	-0.0026	-0.0181	0.0104	-0.0171	-0.0036	-0.0224	0.0124
	(0.0114)	(0.0149)	(0.0111)	(0.0133)	(0.0177)	(0.0263)	(0.0140)	(0.0017)	(0.0202)	(0.0162)	(0.0173)	(0.0174)	(0.0215)	(0.0298)
11	-0.0199	-0.0297	-0.0111	-0.0319	-0.0445	0.0185	0.0136	-0.0022	0.0113	0.0166	0.0010	-0.0264	-0.0560	-0.0487
	(0.0114)	(0.0108)	(0.0099)	(0.0109)	(0.0150)	(0.0261)	(0.0106)	(0.0013)	(0.0229)	(0.0180)	(0.0178)	(0.0213)	(0.0231)	(0.0269)
12	-0.0120	-0.0319	-0.0196	-0.0502	-0.0570	-0.0161	0.0009	-0.0015	-0.0006	0.0114	0.0030	-0.0217	-0.0472	-0.0409
	(0.0097)	(0.0112)	(0.0080)	(0.0098)	(0.0168)	(0.0198)	(0.0103)	(0.0013)	(0.0200)	(0.0172)	(0.0181)	(0.0189)	(0.0194)	(0.0256)
13	-0.0130	-0.0260	-0.0293	-0.0379	-0.0522	-0.0247	0.0170	0.0002	0.0135	0.0381	0.0143	-0.0281	-0.0637	-0.0732
	(0.0069)	(0.0082)	(0.0062)	(0.0072)	(0.0095)	(0.0165)	(0.0072)	(0.0010)	(0.0142)	(0.0123)	(0.0122)	(0.0127)	(0.0147)	(0.0154)
14	-0.0046	-0.0204	-0.0004	-0.0322	-0.0728	0.0044	0.0241	-0.0025	0.0027	0.0392	-0.0333	-0.0196	-0.0391	-0.0362
	(0.0114)	(0.0132)	(0.0101)	(0.0127)	(0.0134)	(0.0227)	(0.0122)	(0.0015)	(0.0209)	(0.0158)	(0.0190)	(0.0192)	(0.0234)	(0.0249)
15	-0.0176	-0.0158	-0.0086	-0.0232	-0.0344	0.0218	0.0278	-0.0010	0.0170	0.0038	0.0172	-0.0229	-0.0527	-0.0708
	(0.0087)	(0.0100)	(0.0077)	(0.0083)	(0.0134)	(0.0174)	(0.0094)	(0.0011)	(0.0178)	(0.0215)	(0.0174)	(0.0148)	(0.0178)	(0.0197)
16	-0.0030	-0.0156	-0.0196	-0.0362	-0.0516	-0.0238	0.0152	0.0000	0.0048	0.0334	0.0023	-0.0294	-0.0622	-0.0615
	(0.0088)	(0.0109)	(0.0079)	(0.0109)	(0.0135)	(0.0200)	(0.0094)	(0.0011)	(0.0164)	(0.0164)	(0.0165)	(0.0153)	(0.0163)	(0.0187)
17	-0.0198	-0.0387	-0.0113	-0.0112	-0.0053	-0.0393	0.0258	0.0002	0.0223	0.0335	0.0111	-0.0096	-0.0209	-0.0313
	(0.0068)	(0.0074)	(0.0059)	(0.0063)	(0.0084)	(0.0152)	(0.0058)	(0.0008)	(0.0143)	(0.0111)	(0.0110)	(0.0124)	(0.0120)	(0.0147)
18	-0.0283	-0.0406	-0.0153	-0.0099	0.0193	-0.0426	0.0229	-0.0014	0.0310	0.0581	-0.0314	-0.0189	-0.0179	-0.0189
	(0.0111)	(0.0132)	(0.0102)	(0.0119)	(0.0158)	(0.0181)	(0.0094)	(0.0013)	(0.0214)	(0.0156)	(0.0168)	(0.0194)	(0.0184)	(0.0210)
19	-0.0101	-0.0162	-0.0141	-0.0072	0.0000	-0.0321	0.0155	0.0013	0.0471	0.0138	-0.0162	-0.0431	-0.0397	-0.0784
	(0.0081)	(0.0118)	(0.0096)	(0.0097)	(0.0133)	(0.0157)	(0.0087)	(0.0012)	(0.0152)	(0.0137)	(0.0135)	(0.0125)	(0.0144)	(0.0248)
20	-0.0179	-0.0173	-0.0073	-0.0193	-0.0127	0.0014	0.0119	-0.0025	0.0600	0.0490	-0.0201	-0.0457	-0.0576	-0.0666
	(0.0089)	(0.0134)	(0.0089)	(0.0100)	(0.0138)	(0.0225)	(0.0109)	(0.0009)	(0.0212)	(0.0173)	(0.0216)	(0.0175)	(0.0180)	(0.0212)
21	-0.0179	-0.0359	-0.0118	-0.0159	-0.0122	-0.0207	0.0221	-0.0006	0.0269	0.0171	0.0119	-0.0144	-0.0326	-0.0493
	(0.0060)	(0.0066)	(0.0056)	(0.0058)	(0.0069)	(0.0114)	(0.0064)	(0.0007)	(0.0136)	(0.0106)	(0.0107)	(0.0119)	(0.0118)	(0.0136)
22	-0.0129	-0.0305	-0.0100	-0.0123	0.0173	-0.0216	0.0141	-0.0012	0.0061	0.0540	-0.0177	-0.0105	-0.0150	0.0046
	(0.0104)	(0.0122)	(0.0090)	(0.0109)	(0.0142)	(0.0212)	(0.0102)	(0.0013)	(0.0185)	(0.0180)	(0.0169)	(0.0171)	(0.0173)	(0.0245)
23	-0.0091	-0.0120	-0.0132	-0.0103	-0.0087	-0.0281	0.0069	0.0006	0.0503	0.0339	0.0027	-0.0233	-0.0389	-0.0546
	(0.0092)	(0.0126)	(0.0089)	(0.0100)	(0.0117)	(0.0199)	(0.0093)	(0.0012)	(0.0191)	(0.0160)	(0.0165)	(0.0179)	(0.0187)	(0.0190)
24	-0.0184	-0.0180	-0.0068	-0.0355	-0.0189	0.0112	0.0111	-0.0029	0.0554	0.0474	-0.0333	-0.0320	-0.0580	-0.0461
	(0.0105)	(0.0131)	(0.0102)	(0.0101)	(0.0133)	(0.0185)	(0.0121)	(0.0007)	(0.0201)	(0.0202)	(0.0192)	(0.0186)	(0.0190)	(0.0276)
25	-0.0160	-0.0288	-0.0162	-0.0261	-0.0226	-0.0454	0.0096	-0.0005	0.0016	0.0133	0.0216	0.0042	-0.0159	-0.0336
	(0.0062)	(0.0064)	(0.0056)	(0.0061)	(0.0076)	(0.0115)	(0.0056)	(0.0006)	(0.0111)	(0.0109)	(0.0110)	(0.0096)	(0.0100)	(0.0129)

26	-0.0420	-0.0341	-0.0038	-0.0128	-0.0012	-0.0361	0.0161	-0.0029	0.0206	0.0420	-0.0046	-0.0064	-0.0073	-0.0181
	(0.0104)	(0.0120)	(0.0106)	(0.0108)	(0.0146)	(0.0208)	(0.0110)	(0.0011)	(0.0190)	(0.0158)	(0.0174)	(0.0172)	(0.0170)	(0.0290)
27	-0.0165	-0.0180	-0.0116	-0.0049	-0.0145	-0.0254	0.0110	-0.0013	0.0334	0.0124	0.0054	-0.0178	-0.0306	-0.0525
	(0.0080)	(0.0103)	(0.0077)	(0.0089)	(0.0106)	(0.0154)	(0.0088)	(0.0008)	(0.0159)	(0.0155)	(0.0152)	(0.0139)	(0.0149)	(0.0163)
28	-0.0155	-0.0256	-0.0118	-0.0252	-0.0134	-0.0271	0.0109	-0.0023	0.0134	0.0260	-0.0054	-0.0156	-0.0380	-0.0378
	(0.0085)	(0.0079)	(0.0072)	(0.0076)	(0.0105)	(0.0116)	(0.0082)	(0.0005)	(0.0178)	(0.0194)	(0.0162)	(0.0134)	(0.0137)	(0.0176)
29	-0.0226	-0.0364	-0.0152	-0.0167	-0.0115	-0.0253	0.0136	-0.0008	0.0228	0.0171	0.0328	-0.0105	-0.0215	-0.0360
	(0.0055)	(0.0061)	(0.0052)	(0.0051)	(0.0070)	(0.0091)	(0.0061)	(0.0006)	(0.0111)	(0.0092)	(0.0107)	(0.0105)	(0.0102)	(0.0124)
30	-0.0357	-0.0466	-0.0094	-0.0120	0.0142	-0.0266	0.0167	-0.0021	0.0231	0.0495	-0.0141	-0.0144	-0.0202	-0.0272
	(0.0089)	(0.0103)	(0.0077)	(0.0090)	(0.0115)	(0.0153)	(0.0107)	(0.0009)	(0.0176)	(0.0134)	(0.0153)	(0.0174)	(0.0171)	(0.0255)
31	-0.0239	-0.0262	-0.0120	-0.0051	-0.0111	-0.0192	0.0082	-0.0004	0.0394	0.0208	0.0125	-0.0212	-0.0267	-0.0494
	(0.0071)	(0.0100)	(0.0075)	(0.0083)	(0.0104)	(0.0147)	(0.0085)	(0.0008)	(0.0155)	(0.0126)	(0.0127)	(0.0142)	(0.0153)	(0.0169)
32	-0.0194	-0.0255	-0.0147	-0.0241	-0.0060	-0.0171	0.0132	-0.0019	0.0354	0.0397	-0.0045	-0.0249	-0.0467	-0.0439
	(0.0081)	(0.0093)	(0.0081)	(0.0067)	(0.0105)	(0.0217)	(0.0088)	(0.0009)	(0.0142)	(0.0145)	(0.0130)	(0.0113)	(0.0122)	(0.0161)

[a] Standard deviation based on 200 bootstrap replications for post-selection estimation is reported in brackets below the corresponding estimated coefficient.

Appendix S4.3: Full Results from Post-Selection Estimation on Relevant Socio-economic Factors and Other Factors

Group Index	educ1	occup1	occup2	lnincome	hp	married	us_born	us_m15	mental	rg_sth
1	-0.0098	-0.0087	-0.0213	0.0218	0.0092	0.0115	0.0611	0.0534	0.0692	-0.0010
	(0.0011)[a]	(0.0067)	(0.0085)	(0.0036)	(0.0092)	(0.0057)	(0.0122)	(0.0138)	(0.0181)	(0.0064)
2	-0.0084	-0.0081	0.0145	0.0301	-0.0037	-0.0179	0.1124	0.0564	0.0775	0.0360
	(0.0026)	(0.0148)	(0.0147)	(0.0071)	(0.0177)	(0.0098)	(0.0177)	(0.0207)	(0.0356)	(0.0107)
3	-0.0115	-0.0275	-0.0242	0.0274	0.0177	0.0147	0.1278	0.0456	0.0681	0.0321
	(0.0019)	(0.0123)	(0.0126)	(0.0048)	(0.0094)	(0.0089)	(0.0122)	(0.0126)	(0.0255)	(0.0090)
4	-0.0088	-0.0162	-0.0058	0.0261	-0.0071	0.0052	0.0869	0.0578	0.0748	0.0161
	(0.0019)	(0.0083)	(0.0126)	(0.0036)	(0.0170)	(0.0084)	(0.0130)	(0.0149)	(0.0197)	(0.0071)
5	-0.0107	-0.0247	-0.0301	0.0075	0.0103	-0.0067	0.0949	0.0594	0.0667	-0.0017
	(0.0011)	(0.0068)	(0.0087)	(0.0037)	(0.0072)	(0.0051)	(0.0117)	(0.0131)	(0.0207)	(0.0059)
6	-0.0058	-0.0326	0.0161	0.0115	0.0223	-0.0260	0.1294	0.0514	-0.0091	0.0301
	(0.0023)	(0.0141)	(0.0140)	(0.0061)	(0.0158)	(0.0105)	(0.0167)	(0.0198)	(0.0259)	(0.0099)
7	-0.0095	-0.0341	-0.0274	0.0092	0.0288	-0.0108	0.1354	0.0429	0.0428	0.0187
	(0.0022)	(0.0125)	(0.0115)	(0.0045)	(0.0100)	(0.0082)	(0.0113)	(0.0128)	(0.0341)	(0.0084)
8	-0.0089	-0.0304	-0.0311	0.0109	0.0009	-0.0081	0.1085	0.0661	0.0469	0.0185
	(0.0022)	(0.0105)	(0.0123)	(0.0048)	(0.0142)	(0.0085)	(0.0154)	(0.0191)	(0.0232)	(0.0089)
9	-0.0081	-0.0077	-0.0093	0.0075	0.0197	-0.0071	0.0813	0.0458	0.0662	-0.0054
	(0.0010)	(0.0068)	(0.0084)	(0.0032)	(0.0087)	(0.0055)	(0.0118)	(0.0125)	(0.0233)	(0.0059)
10	-0.0064	-0.0061	0.0095	0.0199	0.0201	-0.0328	0.0915	0.0267	0.0939	0.0266
	(0.0019)	(0.0127)	(0.0115)	(0.0051)	(0.0174)	(0.0092)	(0.0217)	(0.0233)	(0.0452)	(0.0093)
11	-0.0089	-0.0028	0.0065	0.0137	0.0323	-0.0131	0.1469	0.0308	0.0961	0.0358
	(0.0016)	(0.0119)	(0.0130)	(0.0043)	(0.0121)	(0.0088)	(0.0124)	(0.0154)	(0.0444)	(0.0086)
12	-0.0093	-0.0069	0.0289	0.0177	0.0253	-0.0233	0.1008	0.0339	0.0811	0.0037
	(0.0016)	(0.0096)	(0.0137)	(0.0036)	(0.0112)	(0.0075)	(0.0157)	(0.0183)	(0.0290)	(0.0074)
13	-0.0069	-0.0151	-0.0176	0.0117	0.0200	-0.0038	0.0738	0.0498	0.0557	0.0006
	(0.0011)	(0.0069)	(0.0072)	(0.0032)	(0.0074)	(0.0056)	(0.0111)	(0.0119)	(0.0176)	(0.0060)
14	-0.0055	-0.0188	0.0076	0.0230	0.0268	-0.0344	0.1121	0.0556	0.0440	0.0337
	(0.0021)	(0.0118)	(0.0111)	(0.0048)	(0.0126)	(0.0086)	(0.0154)	(0.0182)	(0.0289)	(0.0094)
15	-0.0073	-0.0201	-0.0125	0.0169	0.0369	-0.0095	0.1438	0.0410	0.0673	0.0249
	(0.0020)	(0.0104)	(0.0116)	(0.0037)	(0.0122)	(0.0085)	(0.0125)	(0.0140)	(0.0296)	(0.0072)
16	-0.0076	-0.0119	-0.0062	0.0161	0.0114	-0.0184	0.1017	0.0556	0.0616	0.0137
	(0.0017)	(0.0091)	(0.0103)	(0.0046)	(0.0120)	(0.0077)	(0.0132)	(0.0146)	(0.0206)	(0.0073)
17	-0.0035	-0.0299	-0.0175	0.0266	0.0137	0.0402	0.0303	0.0310	0.0321	0.0019
	(0.0009)	(0.0067)	(0.0081)	(0.0028)	(0.0063)	(0.0056)	(0.0093)	(0.0103)	(0.0246)	(0.0058)
18	-0.0053	-0.0108	-0.0124	0.0355	0.0330	0.0281	0.0655	0.0290	0.0547	0.0369
	(0.0014)	(0.0117)	(0.0122)	(0.0044)	(0.0104)	(0.0089)	(0.0145)	(0.0163)	(0.0189)	(0.0086)
19	-0.0071	-0.0412	-0.0178	0.0339	0.0101	0.0282	0.0603	0.0026	0.0519	0.0234
	(0.0013)	(0.0086)	(0.0119)	(0.0040)	(0.0081)	(0.0067)	(0.0099)	(0.0118)	(0.0190)	(0.0088)
20	-0.0061	-0.0256	-0.0079	0.0241	0.0184	0.0441	0.0506	0.0081	0.0119	0.0020
	(0.0012)	(0.0080)	(0.0144)	(0.0033)	(0.0091)	(0.0094)	(0.0167)	(0.0180)	(0.0333)	(0.0087)
21	-0.0060	-0.0316	-0.0115	0.0132	0.0179	0.0291	0.0627	0.0449	0.0652	0.0175
	(0.0010)	(0.0058)	(0.0073)	(0.0031)	(0.0051)	(0.0050)	(0.0087)	(0.0103)	(0.0180)	(0.0051)
22	-0.0064	-0.0167	-0.0003	0.0170	0.0173	0.0219	0.1037	0.0396	0.0561	0.0232
	(0.0017)	(0.0101)	(0.0120)	(0.0060)	(0.0095)	(0.0089)	(0.0119)	(0.0145)	(0.0253)	(0.0085)
23	-0.0073	-0.0317	-0.0174	0.0138	0.0040	0.0168	0.0907	0.0154	0.0411	0.0151
	(0.0013)	(0.0090)	(0.0114)	(0.0042)	(0.0080)	(0.0071)	(0.0111)	(0.0130)	(0.0219)	(0.0078)
24	-0.0070	-0.0220	-0.0090	0.0015	0.0146	0.0402	0.0944	0.0438	0.0734	0.0086
	(0.0012)	(0.0097)	(0.0126)	(0.0051)	(0.0093)	(0.0097)	(0.0147)	(0.0154)	(0.0324)	(0.0082)
25	-0.0044	-0.0147	-0.0075	0.0110	0.0214	0.0244	0.0715	0.0473	0.0131	0.0053
	(0.0008)	(0.0055)	(0.0072)	(0.0030)	(0.0058)	(0.0051)	(0.0091)	(0.0098)	(0.0208)	(0.0049)
26	-0.0051	-0.0162	0.0069	0.0242	0.0365	0.0107	0.0940	0.0303	0.0185	0.0172
	(0.0015)	(0.0105)	(0.0125)	(0.0053)	(0.0093)	(0.0080)	(0.0125)	(0.0136)	(0.0329)	(0.0086)
27	-0.0076	-0.0264	-0.0093	0.0171	0.0263	0.0140	0.1076	0.0140	0.0595	0.0110
	(0.0012)	(0.0084)	(0.0092)	(0.0040)	(0.0082)	(0.0068)	(0.0102)	(0.0115)	(0.0262)	(0.0062)
28	-0.0057	-0.0197	0.0014	0.0103	0.0259	0.0246	0.0994	0.0404	0.0275	0.0001
	(0.0012)	(0.0090)	(0.0094)	(0.0038)	(0.0085)	(0.0078)	(0.0111)	(0.0120)	(0.0181)	(0.0065)
29	-0.0044	-0.0212	-0.0158	0.0170	0.0191	0.0296	0.0572	0.0436	0.0394	0.0082
	(0.0009)	(0.0055)	(0.0066)	(0.0023)	(0.0052)	(0.0045)	(0.0087)	(0.0101)	(0.0162)	(0.0048)
30	-0.0059	-0.0149	0.0010	0.0275	0.0278	0.0128	0.0894	0.0297	0.0394	0.0269
	(0.0012)	(0.0093)	(0.0109)	(0.0045)	(0.0079)	(0.0075)	(0.0105)	(0.0122)	(0.0232)	(0.0070)
31	-0.0059	-0.0264	-0.0080	0.0224	0.0099	0.0165	0.0977	0.0066	0.0540	0.0158
	(0.0014)	(0.0068)	(0.0081)	(0.0033)	(0.0068)	(0.0059)	(0.0095)	(0.0117)	(0.0191)	(0.0070)
32	-0.0057	-0.0163	-0.0021	0.0133	0.0164	0.0264	0.0825	0.0371	0.0471	0.0051
	(0.0009)	(0.0073)	(0.0094)	(0.0031)	(0.0077)	(0.0071)	(0.0118)	(0.0126)	(0.0216)	(0.0062)

[a] Standard deviation based on 200 bootstrap replications for post-selection estimation is reported in brackets below the corresponding estimated coefficient.

References.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96** 1348–1360.

Li, Q., Ouyang, D. and Racine, J. S. (2013). Categorical semiparametrics varying–coefficient models. *Journal of Applied Econometrics* **28** 551–579.

Wang, H. and Xia, Y. (2009). Shrinkage estimation of the varying coefficient. *Journal of the American Statistical Association* **104** 747—757.

Department of Econometrics
and Business Statistics
Monash University
VIC 3145, Australia
E-mail: Jiti.Gao@monash.edu

Department of Economics
University of Bath
Bath BA2 7JP, UK
E-mail: bp495@bath.ac.uk

Department of Statistics
University of Pittsburgh
Pittsburgh, PA 15260.
E-mail: zren@pitt.edu

Department of Economics
University of Exeter
EX4 4PU, UK
E-mail: x.zhang1@exeter.ac.uk